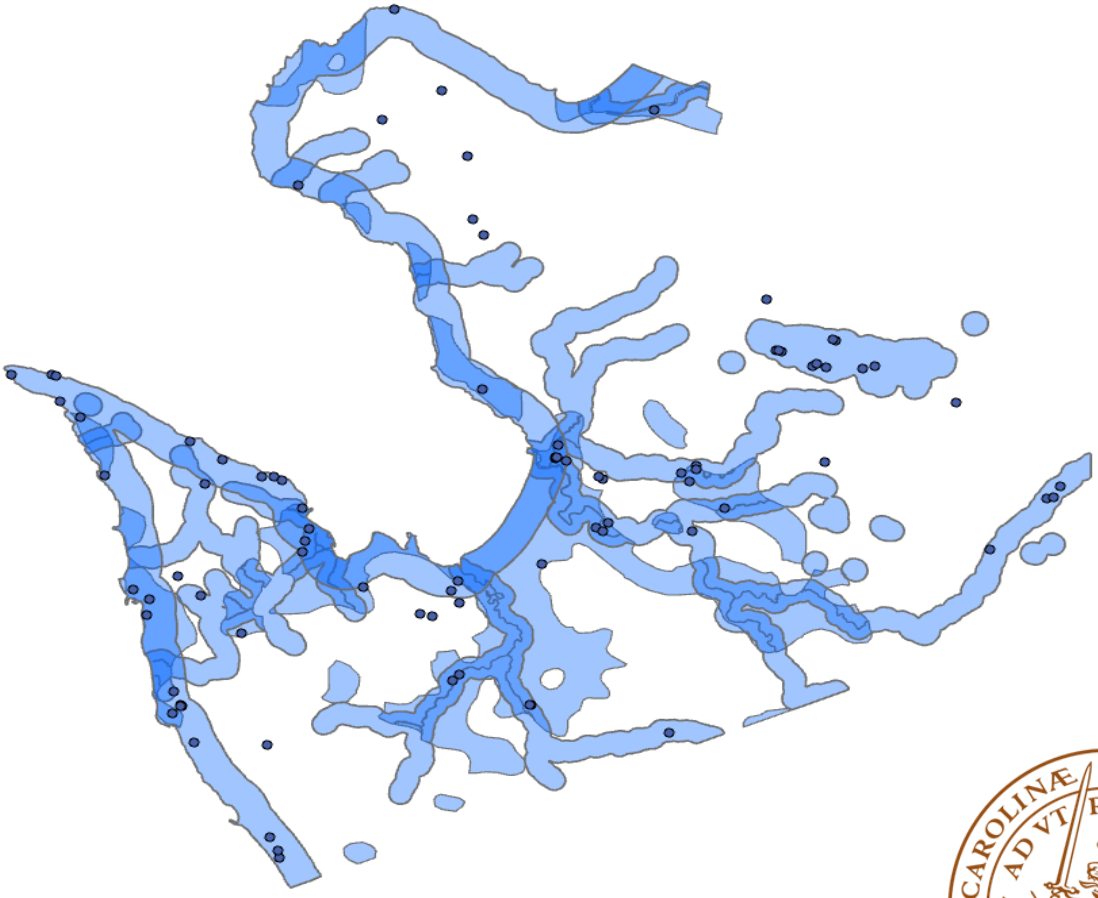
Testing the prehistoric settlement predictors:

A performance evaluation of environmental variables in north-western Scania.



Master's thesis VT 2022
Department of Archaeology and Ancient History
Lund University
Author: André Hjulström
Supervisor: Nicolò Dell'Unto

LUNDS UNIVERSITET

Acknowledgements

First, I want to thank Nicolò Dell'Unto, who has been my supervisor during the planning and writing of this thesis. The process of connecting between the theory and methodology as well as creating a coherent structure in the thesis would have been much more difficult without his help and frequent regular feedback. Secondly, I want to thank Giacomo Landeschi who has given valuable advice regarding the presentation of the methods and the analysis. Lastly, I want to thank the second reader and examiner Fredrik Ekengren, for his feedback on how to further improve the thesis in its final stages of completion.

Abstract

This paper goes over the process of evaluating the environmental variables, which are most likely to predict the locations of prehistoric settlements in a Scanian setting. This is accomplished by selecting variables which have shown to be successful in previous similar study areas and testing the spatial correlation between each variable and known settlement presence. This is followed by an overlay analysis in a GIS environment to test the overlapping areas between the variables. A theoretical background, which legitimise and problematise the methods used are included to put predictive modelling in context with archaeological scientific development. The results show that soil type, distance to major lakes and rivers and distance to coastline shows statistically significant positive correlations with settlement presence. This also holds true for the overlaid intersecting areas between said variables.

Table of contents

1.0. Introduction

Before conducting archaeological surveying or excavations, it is preferable to be aware of the area's most likely to contain the material of interest, not only to save time and funds, but also learn about prehistoric human spatial behaviour. To find out which these areas are, interdisciplinary methods based on geographical theory are often used. The base for these kinds of processes are the environmental variables, which are geographical features whose spatial relation to the studied objects can be measured and mapped. While many geographical features can be related to the presence of settlements, some are more strongly correlated to these objects than others. It is therefore necessary to identify which variables are the most relevant for this task, which is the topic of this thesis.

This will be done by first discussing the background of spatial analysis and predictive modelling in archaeological research. We will identify how these variables are used as the basis for constructing parameters within a GIS environment and later used as components in a statistical analysis to achieve a successful predictive model.
Previous research conducted in study areas similar to the one chosen in this study, is presented and used to identify which variables could be relevant for predicting settlement presence in the chosen study area. This assumes that similarity in geographical features and proximity between areas yields approximately the same correlation patterns. This is a topic which will be further explored later on.
Following this, I will present the theoretical background behind predictive models derived from geographical features in relation to human activity.
The methodological steps necessary to take for assessing the analytical value of the chosen parameters and how these can be made usable for a statistical analysis are described thenceforth. The material and processing methods applied to them are presented in the chapter thereafter.
This is followed by the analysis chapter, where the methodological steps described previously, are applied on a sample of sites within the chosen study area.
Finally, I will discuss the results with the goal of answering the research questions and make suggestions for future research and predictive model construction.

§ 1.1. Purpose and aims of the study

The purpose of this study is to lay the necessary foundation for future research, regarding the creation and validation of an archaeological predictive model for prehistoric settlements in a Scanian and southern Swedish setting. This aim is fulfilled through discussing and testing the correlation between environmental parameters and the location of settlements through inductive methods with already identified sites in the study area. The main aims are therefore to establish which the most prominent environmental variables are and, in the process, further develop our understanding of variable selection for this purpose.

1.2. Research questions

There are three main research questions which needs to be addressed in order to approach the answers which this thesis is intended to provide.

- Can a clustered pattern be observed, which would indicate the existence of influence from one or more environmental parameters?
- Based on the accessible data, which variables are the most important for predicting the presence of prehistoric settlements in Scania? What are their relative levels of importance according to observed results and their performance in conjunction with each other?
- Is it possible to extract fundamental principles and notions from this conclusion, which might aid the development of predictive models in other regions?

1.3. Introduction to predictive modelling

In this chapter, the fundamental principles behind archaeological predictive modelling and a brief research history will be discussed in three different aspects: Spatial analysis, GIS technology and statistical techniques. The first part 1.3.1, which is addressing the topic of spatial analysis, is covering the relationship between distance and geographical location with human settlement distribution and how these fits into the subject of archaeology. The second part, subchapter 1.3.2, covers the role of GIS technology in this matter. The third part 1.3.3, covers the different statistical approaches that are used to describe and analyse these relationships. Modern archaeological predictive modelling (APM) consists of these three parts, which makes them important to address and explore in order to understand the practice of APM.

1.3.1. Spatial analysis and predictive modelling

Studying the spatial relationships between archaeological material and their surroundings has always been of interest to archaeological researchers, it gives us a better insight into prehistoric human activity by giving the objects of study a geographical and spatial context to relate to.
As described by Stanton W.Green: “To a certain extent, archaeology can be viewed as a discipline involved in sampling space in order to understand human behaviour” (Allen & Stanton. 1990, p.3).

An example of early spatial analysis in archaeology is the work conducted by the 19th century British archaeologist General Pitt Rivers, who documented multiple scaled plans and sections used to display the locations of artefacts and features on a site in three dimensions (Wheatley & Gillings 2002 p.2-4).

From the understanding of spatial relationships, we can draw several conclusions, ranging from prehistoric human exploitation of the environment’s natural resources to social power dynamics within a single settlement.

Based on the relationships between human activity and geography, we can construct models that explain and visualise them in the shape of maps and tables.
If a significant correlation can be observed between signs of human activity and environmental variables, we can make predictions about the probability of finding archaeological material in areas that have yet to be surveyed. This leads to the subject of archaeological predictive modelling (APM).

A predictive model is commonly described as a technique to predict the location of archaeological sites or materials in a region based on a sample or fundamental notions concerning human behaviour (Verhagen 2018, p.1). This definition captures the essence of what a predictive model is, i.e., a tool, whose reliability depends on our current knowledge of prehistoric human behaviour, as well as the quantity and quality of the data sample we have access to. A predictive model should not be treated as a truth-teller which exceeds human capability, as it is entirely reliant on our ability to recognise the relevant components that the model should consist of.

While settlement patterns in an archaeological context have been studied since the middle of the 1900's, the proposed links between human activity and environmental variables were largely anecdotal. The patterns were observed and pointed out, but no real efforts to map the statistical correlation between the environmental parameters and signs of human activity of this pattern was made until the 1970's, when data analytics was introduced to archaeological settlement studies (Judge & Sebastian et al 1988, pp.30-32). These studies were nevertheless essential in providing the theoretical framework through which predictive models would later be developed.

Among the pioneers of archaeological predictive modelling, the most prominent figure is Kenneth L. Kvamme, who has made several contributions which has led to the fundamental groundwork for predictive modelling within archaeology. The most notable of which is the chapter "Development and testing of quantitative models" within the book "Quantifying the present and predicting the past" of which Kvamme contributed to. In this chapter, he establishes the fundamental expression for the relationship between site presence and geographical extent all within the quotient named after him "Kvamme's gain" (Judge & Sebastian et al 1988 p. 329).

As described by another distinguished APM researcher, Martijn Van Leusen, the development of archaeological predictive modelling has two main root causes: One is its utility in the early stages of spatial planning for major projects and the other is the scientific role it has in developing our "insight into past human behaviour in relation to the landscape".

The method of filtering out potential areas of interest containing cultural remains became necessary when CRM (cultural resource management) developed in the USA in response to legislation regarding the protection of cultural remains. This development would also be seen in Europe, when a meeting was held by the Council of Ministers of Europe in Valetta, Malta in 1992. In this meeting, the representatives of the member states came to an agreement (known as the Malta agreement), which requires the member states to devise means for protecting their cultural heritage as part of the regular spatial planning process (Van Leusen 2005, p.1).

Predictive models have been seen as problematic though, due to the potential exclusion of areas containing sites. In cultural heritage management, this is generally seen as unacceptable, but in archaeological science it is seen as a chance to further develop the accuracy of the models.

Computer technology developed in the early 1980's to the point where both digital processing and visualisation became possible to the average researcher without a data scientific or computer scientific background. Archaeologists were early in adopting this technology which would be called Geographical information systems (GIS) (Verhagen 2018 pp.1-3). This will be further explored in the next part.

1.3.2. The impact of GIS technology

During the processual archaeology movement beginning in the 1960's, external factors were emphasised to be the key influence behind human behaviour. This behaviour leaves patterns in space, which in turn can be measured and quantified to identify the generative process. This approach came to be through the applications of a wide range of spatial analytic methods and techniques (Wheatley & Gillings 2002, p.6).

In conjunction with this new approach to spatial analysis of quantitative data, the emergence of Geographical information systems technology created new possibilities never seen within archaeological spatial analysis. The main difference between GIS and computer aided design and mapping programs is the spatial database, which can contain a large amount of information that is spatially referenced and can be subject to queries and analysis (Wheatley & Gillings 2002, p.11). This is very useful in archaeological practice and research, where often large amounts of data are handled, and spatial references are needed.

While Geographical information systems have played a major role in archaeological practice such as field walking and distribution maps, it has also played a very important role in the development of predictive models. A GIS provides the confines in which we are constructing the environmental parameters based on geographical and spatial data. The quality of a predictive model is not only dependent on the mathematical principles behind it or the logic applied, it is also dependent on the quality of data and the ability to process it through the software. What determines the quality of the data is mainly twofold (Chapman 2006, p. 54):

- Degree of resolution, which mostly refers to the size of the image pixels from photographs. It could also mean the level of detail in surfaces which are built based on relative numerical values, such as elevation surfaces.
- The positional accuracy, which refers to how well the data is defined to be positioned compared to its actual position. This is crucial for any type of measurements to be done based on the parameters. The accuracy is affecting their assumed extent and relative position to each datapoint.

The importance of these two factors when using or creating environmental parameters within a GIS cannot be overstated. The ability to create reliable datasets which will be the basis for further processing in a GIS environment depends both on the instruments used in the field and on the level of competence of the surveyor using the instruments (Chapman 2006, p.54**)**. Limitations in both should be addressed when creating predictive models based on processed datasets derived from field surveying.

The ability within a GIS environment, to process geographical data into new datasets which could function as environmental parameters, along with the ability to handle a large amount of data with

spatial references, is invaluable to the development of predictive modelling. Also, the simplicity of a well-made predictive map is a great tool for convincing local governments and developers of the archaeological potential of an area and thus prevent the risk of damaging sites (Verhagen 2007, pp. 17-18).

Without a GIS, the construction and visualisation of a predictive model would be done manually in the field or drawn on a map. This would be exhausting and time-consuming with potential errors made through mistakes. These issues can be eliminated in a GIS, where distances and measurements of the variables are done by the software. Assuming that the input data is accurately measured, and an adequate resolution is acquired through satellite, aerial and/or terrestrial instruments, this will enable the task to be done faster, easier and more cost effective (Allen & Stanton 1990 p.165).

Within a GIS, it is possible to reconstruct landscapes based on elevation data acquired through measurements done from aerial or terrestrial vehicles. The output of this process is called a digital elevation model or a DEM, which is a surface displaying the distribution of elevation values over an area through attribute values in cells. These surfaces are often displayed by a grid of squares with individual elevation values, whose size depend on the spatial resolution. They could also be displayed as TIN surfaces, which represents the topography of the landscape through interconnected triangles (Allen & Stanton 1990, pp. 166-167).

An important aspect of the DEM's for archaeological predictive modelling is the ability to create several different datasets from it. A generated surface displaying elevation or slope values, can be the basis for creating cost-surfaces, where the pedestrian traversing difficulties are visualised. Another example of surfaces created based on elevation are viewsheds, where the area visible from a given point is displayed (Chapman 2006, pp. 22-23). The utility of a DEM for constructing environmental parameters has been tested and verified through several studies. Perhaps the most widely used of whom, for its utility, is slope. Slope is calculated in the GIS software through interpolating the surface discrepancy between the individual cells with elevation attributes. This makes it an estimate rather than an exact description of reality, due to the dependency on the level of resolution, which decides how detailed and accurate the interpolated surface is (Chapman 2006, p. 117).

The importance of the slope variable for predicting settlement presence seems to be different depending on other environmental variables that are present. Steep slopes seem more tolerable when there is a close proximity to water sources, while mild slopes are favourable close to smaller water streams, where there's a lesser risk of flooding. This is presumably due to the risk of flooding when close to major rivers at low elevation. A study suggested for example, that areas vulnerable to flooding had a significant underrepresentation of Neolithic settlements, despite them being located in otherwise favourable locations (Mihu-Pintilie & Nicu 2019, p.14).

This difference of importance depending on other external factors seems to also be the case for the viewshed. The less obstructions and the higher elevation, the more visible a point is, which at the same time grants greater visibility from said place. The hypothesis is then naturally that high visibility is preferable to low.

This doesn't seem to always be the case on all places, which was demonstrated in a study presented in an article named "*Using Viewshed Analysis to Explore Settlement Choice: A Case Study of the*

Onondaga Iroquois". In this study, Iroquois settlement location choices were studied with respect to a DEM-derived viewshed. The author of the article recognised that none of the settlements within the study area had a "commanding view of the landscape", instead preferring low hills and gentle slopes. The reason for this is interpreted to be longer growing seasons, sacrificing a naturally more defensive position for agricultural convenience (Jones 2006, p. 14). This example shows that viewsheds can be useful but should be taken into account with other environmental parameters when predicting settlement locations.

In summary, archaeological predictive modelling has been greatly facilitated by the implementation of GIS technology. It has enabled archaeological practitioners and researchers to create basic models rather easily by utilising the large storage capacity of geographically referenced data in the geodatabase, the convenient visualisation that the GIS software provides and the ability to create parameters based on geographical data. Although it has undoubtedly made the process easier, all the limitations of the software will also be what limits which methods and analyses we are able to perform in that environment if we are solely reliant on a particular software. These methods and analyses will be further explored in the following chapter.

1.3.3. Statistics in archaeological predictive modelling:

To perform a spatial analysis, the features and objects we wish to study need to be defined in quantitative terms. These terms are expressed as positions, sizes and shapes or the boundaries of said objects. These objects are our data points, which are systematic observations done by professionals. The systematic nature of the measurements and observations make their definition unambiguous, allowing us to have a reliant standard of measurement (VanPool & Leonard 2011, p. 6). This means, in geospatial predictive modelling terms, that we have a consistent way of categorising sites. A standard projected coordinate system also enables our model to be accurate when applied to geographically referenced data points.

If we wish to study the environmental variables as confined spaces which contain the objects whose presence we want to predict, we need to treat said space and objects as data, not actual physical geographical areas and sites. This simplification of reality is necessary to erase any ambiguity which otherwise may demand too many variables to be considered than what is possible or available.

Within the discipline of archaeology, we are reliant on incomplete data due to the inability to retrieve and analyse everything that has taken place on any given site in the past. This incompleteness makes it necessary to acquire a sample of the features or objects we want to research (VanPool & Leonard 2011, p. 2). We will have to assume that this sample is an adequate representation of the objects or features in terms of distribution, qualitative attributes, and quantity in order to generalise the observed pattern to the entire area of interest. The reliability of this will be further explored in the analysis chapter.

At its core, predictive modelling is a statistical evaluation of the probability of a feature being present within a confined space, which is the dependent variable. The value of the dependent variable depends on one or several independent variables, which in this case are the environmental parameters.

There are two ways that the values of individual independent variables can affect the dependent variable (Verhagen 2007, p. 74):

- Negative correlation: The higher the value of the independent variable is, the lower the value in the dependent variable is. An example of this may be distance to water sources and slope gradients, which often negatively correlates with the presence of remains from human activity
- Positive correlation: As opposed to negative correlation, the higher the value of an independent variable is, the higher the value of the dependent variable is. Examples of this can be the amount of exposure to sunlight an area has or the total area which is visible from any given point within a zone.

The level of correlation is represented by a scale ranging from the values -1 to 1. With -1 being a perfect negative correlation between the dependent variable and the independent variable, i.e., if one increases the other one is decreasing to the same degree. The opposite relationship is true if the level is 1, then an increase or decrease in the value of one variable will also result in the exact same change in the value of the other. A level value of 0 represents a perfectly random relationship between the independent and dependent variables, which means that there is no correlation.

There are multiple different ways to create the probability distribution zones within a model, the most common of which is multiple logistic regression analysis, which is a statistical technique that expresses the relationship between individual independent variables and the dependent variable separately (Drennan 2009, p.264). Or binomial logistic regression, where there are only two possible alternatives (Yaworsky et al 2020): either sites are expressed to be present within the area of the model or it isn't (The model consists of "yes" and "no" zones). This lack of nuance isn't a problem, because of the dualistic nature of the issue: Either a model predicts a site presence, or it doesn't.

When expressing the probability of site presence and absence in a formula, this is how it's usually done (Ejstrud 2004, p.10):

$P(site \mid A_x B_x C_x) = 1\text{-}P(\sim site \mid \sim A_x \sim B_x \sim C_x)$

Where P stands for probability, site stands for the feature whose occurrence we want to examine, "|" means "given that ". A_x, B_x and C_x are the independent environmental variables and "~" means "not". The statement of the formula therefore reads out:

> *The probability of site presence given the following parameters equals the difference between 1 and the probability of site absence, given the absence of the parameters in question*

This statement assumes that we have all the necessary information about the location of all possible sites in the study area, leaving no room for uncertainty. Although when we are analysing a distribution of sites within a geographical area, we are dealing with incomplete datasets, due to the inability to survey an area completely in most cases. This results in areas being absent of sites, which

could mean that they are not yet surveyed, not necessarily avoided by ancient people for settlement. Therefore, the more well surveyed an area is and the more data points we have, the more accurate a picture we get of the performance of the parameters when we test them against the known sites.

There are several different techniques, which has been implemented in APM, the most common of which will be represented below:

Logistic regression is a technique, which is non-parametric (Svedjemo 2003, p. 8). This means it's unassuming of both the nature of the variables (if they are nominal or expressed in a ratio) and how the data is spatially distributed. This allows us to use non-quantified variables such as soil type along with quantified variables such as distance from the coastline with the same technique, which is needed if we want to compare predictive power between variables.

Another statistical technique, which has seen recent use within archaeological quantitative analyses is Bayesian statistics. While it has been used primarily within radiocarbon dating, it shows some promise for predictive modelling as well, where it has started to find its way into archaeological predictive modelling (Otarillo-Castilla & Torquato 2018, pp. 8-12). The technique builds upon our previous knowledge of the probability of a hypothesis being true or false depending on what data we have acquired as of yet and what that tells us.

What bayesian statistics try to solve is the following:

"Given a certain outcome, what is the probability that the parameter or parameters are behind the result?" (Ortman et al 2007).

The probability of our hypothesis being true changes depending on new input when data is collected from the surveying process, which refines the model. This is very useful for evaluating the predictive power of individual parameters, with a hypothesis made of them containing sites to a higher degree than random chance.

After the model is constructed, its performance is expressed by a value, which is called Kvamme's gain. The formula is as follows (Judge & Sebastian et al 1988, p. 329):

"*1 – (Percentage of total area covered by the model) / (Percentage of total sites within model area)*"

As discussed previously, the closer the gain value is to 1, the better the predictive model is in narrowing down the areas which most probably contain sites, with the ideal predictive model of a Kvamme's gain value of 1 covering a small percent of the entire study area and includes all of the archaeological sites.

Another very important part to consider is data representativity. If we rely on a predictive model, which tells us where it is worthwhile to survey and excavate, we might neglect surveying the areas where low probability of finding sites is suggested. This could have the consequence of us further reinforcing the result of the model by finding more sites within the high probability areas than the other areas, making the model a self-fulfilling prophecy rather than an accurate tool to predict the distribution of sites in relation to environmental factors. It would therefore be more statistically rigorous to survey the entire study area equally (Verhagen et al. 2010).

Often overlooked, but very relevant when discussing sample and study area size is statistical significance, i.e., the probability of the observed pattern to be that of random chance. The importance of this will be explained later in the methodology chapter of this paper.

In essence: The chosen statistical technique will define the shape and extent of the surfaces we expect will contain sites within our study area, which can be as important as the choice of environmental parameters for our model. After the environmental variables have been chosen and the analysis made, the model is tested through calculating Kvamme's gain, which will give us back an assessment of the model's performance. The model is then validated through applying it to other study areas, which will either reinforce its usability or put it to question.

1.4. Previous research

In this chapter, different case studies are presented to illuminate which kinds of variables we can expect to be relevant to use in a Scanian environment to predict the locations of prehistoric settlements.

Because every landscape and environment is different, we have to evaluate which parameters to assess based on the local geographical features. In this paper, I will investigate which variables are the most relevant for predicting prehistoric settlement locations in a southern Swedish setting with north-western Scania as study area. To get a good starting point in this endeavour, I will go through six different case studies, all of which relate to the chosen study area in geographical proximity, extent, and environmental characteristics. The variables used in these studies will form the starting point from which the variable selection will occur.

1.4.1 Brandenburg, Germany:

In 2010, Benjamin Ducke wrote the paper "*Regional Scale Predictive Modelling in North-Eastern Germany"*, in which he describes the design, implementation and application of a region-wide predictive model of prehistoric settlements and graves in Brandenburg state, North-eastern Germany. The expressed goal was to provide the necessary information and research perspective to understand the archaeological landscape of Brandenburg. (Ducke 2010, p.1)

The environmental parameters which were included are:

- Soil type
- Elevation, slope, aspect
- Rivers and lakes
- Previously identified graves and settlements within the study area

The Euclidean distance to water bodies such as lakes and rivers as well as sites were of interest, as such the parameter was defined as a buffeted zone around these items, within which an overrepresentation of sites are expected to be found. By experimenting with the data, Ducke found that a buffer zone of 500 metres was sufficient for all kinds of water bodies, although a differentiation of distances based on the type of stream improves model quality he adds (Ducke 2010 p. 2).

The most promising soil type was regarded to be clay and peaty soils, due to the fertility of the soil. However, the importance of this type seems to depend on time-period, with a high amount of Neolithic sites showing preference for it, while sites from the bronze age instead showing a preference for a high altitude with good visibility over the landscape (Ducke 2010, pp. 4-5).

Ducke concludes through testing that terrain curvature was not relevant and thus excluded from the analysis, although the data suggests that there seems to be a slight preference for surfaces where there is an east and south-east facing aspect. He also added buffer zones with a radius of 500 metres from a randomised sample of the known sites.

The study area was divided into 13 different sections based on geographical boundaries which has divided the region historically, in order to test the model on areas with different topographical characteristics and archaeological contexts. The statistical technique used is Dempster-Shafer theory of evidence. The author motivates this choice based on the presence of uncertainty, which techniques such as Bayesian statistics and logistic regression do not account for but is relevant for estimating site presence probability in cases where absence of one parameter and presence of another might contradict each other (Ducke 2010, pp. 2-3).

The results showed that soil quality and proximity to rivers were very significant for predicting settlement presence (Ducke 2010, p. 4). The presence of other sites was also an important factor, with a radius of 2 kilometres around each site, the total area of all buffer zones combined an area of 10% of the study area, while containing almost half the known sites (Ducke 2010, p. 3), this gives a Kvamme's gain value of between 0.75 and 0.8, which makes the parameter very useful for predicting settlement presence.

The region of Brandenburg is close to Scania, with similar geographical characteristics. This fact along with the promising results that this paper shows, one could make a convincing argument that it could and maybe even should be used as a reference when evaluating the possibilities of predicting site presence in southern Sweden. With that said, the spatial and temporal scales are very large, which makes it hard to pinpoint which variables perform the best in predicting archaeological settlements or other sites in general, because of the seeming scale-dependence of this matter.

1.4.2 Eastern Jutland, Denmark:

In the academic paper "*Ejstrud, Bo "Indicative models in landscape management: Testing the methods", 2003"*, the archaeological researcher Bo Ejstrud wrote a chapter titled "*Indicative models in landscape management: Testing the methods"*.

In this chapter, he discusses and compares the different methods of producing archaeological models for locating sites, which are invisible on the surface and can thus be ignored and damaged by agricultural or infrastructural projects in Denmark.
The purpose of this is to conclude which type of technique performs the best on the archaeological material within the study area in eastern Jutland (Ejstrud 2003, pp. 1-2).

The environmental parameters Ejstrud selected when creating the models were:

- Soil type
- Slope, aspect, relief, and exposure (all derived from a digital elevation model)
- Distance to water
- Presence of wetlands

The author judges the aspect variable to be of "absolutely no importance to settlement location in the area investigated, at any time in history". He adds though that many different pieces of information can be derived from a DEM, which could be useful. Variables assigned to altitude and altitude discrepancies are expected to have a weak performance in a flat landscape such as in Denmark (Ejstrud 2003, p. 3).

The author mentions that Mesolithic graves are located on the same site as contemporary settlements, but in later periods they are located on ground with high elevation for visibility (Ejstrud 2003, p. 6). This information may be useful if we want to identify the potential location of sites based on the presence of Mesolithic graves.

For maximising the validity of the models, the author restricted the data points from 10,000 observations to 1000, all of which were dated and had their positions adequately defined within the landscape. The temporal scale was not an issue, due to the well-defined periodical divisions between the 1000 data points, although the author deemed it too time-consuming to make separate models for each time-period. He therefore grouped all material from the Mesolithic to the early roman iron age (9000 BC to 200 AD) together (Ejstrud 2003, p. 5).

The author describes a settlement pattern where the early settlements during the Mesolithic were always close to water sources such as rivers, lakes, and coasts, with a gradual shift to the inland due to the need for access to grazing pastures and fertile soil for agriculture, while still being close to water sources. In the late Bronze age, the settlements are observed to be located out on the moraine plains, instead of being close to water streams and coasts (Ejstrud 2003, p. 6).

The different methods tested were the following (Ejstrud 2003, pp. 7-11):

Boolean overlay/Binary addition: The study area is divided into zones where there either is an expected site presence or where there isn't one. All variables are, for simplicity's sake, deemed to be of equal importance as each other. The more positive variables overlapping, the more probable site presence is estimated to be.

Weighted binary addition: This method works the same as the one mentioned above, but the variables are assigned weights based on how frequently sites occur in them individually. The more the site presence frequency is affected by a decrease or increase in value of a variable (for example distance to water sources), the more weight the variable is assigned to have. This method cannot be used on variables with a nominal scale, like soil types for example.

Logistic regression: This method measures the probability of site presence or absence as a function of variable presence. When applied to a study area containing previously identified sites, it assumes that areas (often raster cells in a GIS) representing "non-sites", i.e., where no currently known sites

are located, are areas with a negative output and therefore do not contain sites. Logistic regression also assumes that the variables are normally distributed (the mean values are the most common values).

- Dempster-Shafer theory: To solve the issue of assuming that areas representing non-sites are areas actually absent of sites, an element of uncertainty could be accounted for. This is a necessary step for accurately assessing the potential of an area containing an incomplete set of data, which most archaeological study areas are. This uncertainty, accounting for ignorance, is built into the method which the Dempster-Shafer theory is based on. The concept of absence is different from the non-sites of the logistic regression method, in this case, it represents areas where surveying has been avoided rather than areas where no sites exist, the resulting zones are thus not treated as negative evidence for potential site presence.

The result of Ejstrud's comparison shows that the Dempster-Shafer method performs better than the other methods, especially when it comes to settlement prediction, with a Kvamme's gain of 0,83 for mesolithic, 0,48 for neolithic and 0,33 for late bronze age and early iron age (Ejstrud 2003 p. 11). A major downside of using Dempster-Shafer theory is that it is relatively complicated to use and isn't as supported in commonly used GIS software's as compared to the other methods.

Ejstrud's work is an excellent reference for selecting appropriate environmental parameters for Scania, due to historical connections between regions and very similar geographical characteristics. His work also provides a valuable insight into which statistical methods might be the most appropriate to use based on the given variables.

1.4.3 Archaeological predictive modelling in Sweden

Due to the low number of archaeological predictive models in Sweden produced within academia as well as cultural heritage management, it is important to carry out a study with the goal of assessing which geographical parameters may be important to consider when the construction of these models may become relevant. There are however noteworthy contributions on this topic:

The doctoral thesis "*Löwenborg, Daniel "Excavating the Digital Landscape: GIS analyses of social relations in central Sweden in the 1st millennium AD". 2010*" includes a paper written by the same author, which evaluates the potential of predicting burial mound locations in the region of Västmanland in central Sweden. When constructing the model, he considers variables such as soil type, degree of fragmentation in the landscape (the more fragmented/presence of impediments, the more difficult agriculture is), distance to water bodies, survey (density of known registered sites) and topographic elevation. The last variable was deemed not useful due to the flat terrain in the area, although the author concluded that it was useful in reconstructing the prehistoric shoreline (Löwenborg 2010). The similarities between the distribution of burial and settlement sites seems to vary depending on time-period, displaying a gradual change from a low to a high position over the course of the Mesolithic to the early iron age as seen in the study done by Ejstrud mentioned previously (Ejstrud 2004 p. 6). Due to the flat landscape of the study area, one could assume that the degree of difference between settlement and burial mound distribution would in this case be low,

which could entail that the mentioned variables in the study done by Löwenborg also would be appropriate to predict the presence of settlements.

In the master's thesis "*Asserstam, Marcus. Predicting mesolithic pioneer settlements in Eastern Middle Sweden, 2010.",* the author is constructing a model based on geographical parameters for predicting settlement locations in a eastern Swedish setting. The chosen parameters are slope degree, distance to coast and soil type. All of these are chosen based on assumptions regarding preferences in habitability and proximity to lanes of transportation (Asserstam 2015)

The paper "Predictive models for iron age settlements on Gotland 200-600 AD " was written in 2003 by the author Gustaf Svedjemo. The author is describing an implementation of a predictive model which he himself created for predicting the locations of iron age settlements on the island of Gotland. Like in the doctoral thesis by Daniel Löwenborg, the author deemed elevation to be lacking importance due to the flat landscape of Gotland. Among the variables used are like the previous examples soil type, but also historical land use and settlement patterns from the 18th century, information of which the author retrieved from historical maps (Svedjemo 2003).

The successful implementation of archaeological predictive modelling methods on Swedish material shows that it is indeed possible to create predictive models in this setting. All of these case studies have been done within the frames of rather small study areas, which is appropriate considering the research questions, but makes them rather region-specific. For this reason, these case studies should be treated as guidelines for variable selection but be critically evaluated for their relevance in the specific study area we wish to study, which in this case is Scania. The more predictive models that are constructed within a large area, even if the individual study areas of the different models are small, the more knowledge we gain of which variables are the most relevant for the region at large

1.4.4 The importance of case studies

While evaluating which variables are the most useful for predicting prehistoric settlements, we should not reinvent the wheel where it's not necessary. It's more appropriate to build further on the accumulated research that's been done and conclusions that's been proven successful. Scientific validation is another reason why previous research is important to have as a guideline of variable selection. If the choice of variables is based on pure assumptions without scientific ground and the result of the model is successful without a good understanding of the reason behind why, then the model may not be replicable in another environment, making the model practically useless as an explanatory tool.

If the study area is widely dissimilar to ours in terms of e.g., elevation variance, then this variable will most likely not have a similar role in our study area as in the previous project. Depending on which cultures and what time-period we are studying, we can expect settlement patterns to be different, with the reasons behind this difference not necessarily being purely environmental. In order to conduct a spatial and statistical analysis in a predictive model based on geographical data, we need to make sure that these are quantifiable, i.e., are able to be expressed in numerical values within a scale. Alternatively, that we can use them as Booleans, which means they either excludes all areas outside or within them.

From the case studies provided above, we can observe that the environmental parameters are chosen based on conventional notions regarding settlement patterns relating to these variables in and around the study area. This is generalised across a wide timespan of several thousand years in most cases, due to the issue of temporal resolution.

In the study done by Ejstrud (Ejstrud 2003), dividing the data based on chronological era was shown to be useful in highlighting the varying importance of the environmental parameters in relation to the time period of which the settlements belong. The reason why this usually isn't carried out when predictive models are created in general, is probably because of the small amount of data points we often have access to and therefore aren't able to use as a base for a statistical analysis.

The study areas from the case studies presented here are close to each other both geographically and in how they resemble each other, this gives us a good reason to assume that the parameters were chosen based on this information. Additionally, it is possible to make an argument of why these same parameters should be hypothesised to be of importance in Scania as well due to geographical similarities and historical links to these regions which might indicate a similar settlement pattern being present.

There are several different factors that influence the ability to be consistent when implementing the same methods and testing the same variables as done in other case studies. Data quality in terms of spatial resolution and measurement accuracy will most likely always vary between regions and the time the studies were conducted. The constructed parameters will also vary a lot depending on the decisions made by the individual researcher in terms of extent, accuracy, and rigorous implementation of the methods.

Therefore, there will most likely never be an exact methodological framework, which is applicable to every area regarding archaeological predictive modelling. With that being said, it is important to take inspiration from previously carried out research projects in order to further enhance the practice of scientific APM.

The notions, which we base the methodology on are derived from the theoretical development which has been taking place within academic archaeology and related disciplines. The following chapter will go over this topic.

2.0. Theory

In this chapter I will go through the theoretical frameworks, which influences archaeological predictive modelling relating to geography and space, along with some of the most prominent theories behind prehistoric human locational behaviour and settlement patterns. This is followed by presenting the two main ways of approaching the construction of archaeological predictive models and the importance of the theoretical frameworks in the creation process.

More specifically, I will first present the paradigm shift that has been proposed to be taking place within archaeology by Kristian Kristiansen (Kristiansen 2014) and how predictive modelling fits into that shift. Secondly, I will discuss the issue of human experience and decision making as environmental variables for predictive modelling, as presented by the researchers Lock, Kormann and Pouncett (Lock, Kormann & Pouncett 2014) and Gillings (Gillings 2012). Thirdly, the

importance of measuring the degree to which a distribution pattern can be considered as clustered is discussed, with Tobler's first law of geography in mind. Lastly, inductive, and deductive modelling is presented and discussed, namely what separates the two categories and what place and roles they both have separately from and with each other.

2.1. The new paradigm

For the last four decades there has existed a consensus among the archaeological academic community regarding the way we are expected to approach the way we are interpreting the past through the archaeological material known as the paradigm of post-processualism. This paradigm was characterised by the renunciation of the methods and theoretical frameworks imported from the natural sciences by the archaeological processualists, which according to the post-processualists is a dehumanisation of the past (Kristiansen 2014, p. 12).

Recent developments in genetic research have enabled archaeological researchers to make new interpretations about human migration and origins by analysing genomic data which previously weren't possible with only mitochondrial DNA. This development along with extensive isotope analysis of metals and our reliance on databases due to the acquisition and usage of large amounts of data, have spearheaded archaeological research in a direction where natural scientific methods are yet again in the centre. While the post-processual framework is "withering" as described by Bjørnar Olsen (Kristiansen 2014, pp. 13-14), it is still very relevant within academic research and is contending with as well as working in conjunction with the new quantitative leaps**.**

This "third science revolution" proceeding the second, where radiocarbon dating was introduced, have paved the way for other quantitative measures of interpreting the past, such as palaeobotanical reconstruction of past landscapes and digital models of settlements based on geographical and archaeological data (Kristiansen 2014, p. 18).
As a means of visualising this development, Kristiansen has constructed a model which resembles a wheel representing the theme of mobility, which he describes as "the main research theme during the next two decades". Mobility means the study of all things moveable, i.e., humans, animals, raw material etc…
The spokes of the wheel are represented by the methods through which we are analysing and theorising about mobility. The opposite facing spokes represent two different theoretical or methodological approaches to the same theme, for example the dichotomy of genetics & heredity and culture flow, which are two different approaches to the question of migration and how this can be traced in the archaeological material. In the same way, human activity and mobility can be studied through the dichotomy of settlement and landscape modelling (Kristiansen 2014, pp. 20-21).

By establishing theoretical frameworks regarding the relationship between landscape and settlement, we can also enhance our understanding of the environmental parameters which predict tendencies of human settlement.

2.2. Human experience as variables

Environmental variables do not exclusively serve as natural resources or convenient surfaces to build on or to cultivate for human settlers, they can also represent the potential and limitations of human senses and mobility. One example of this is a study, where the researchers presented several different potential routes between barrows in a landscape, depending on cost of movement in terms of euclidean distance and slope, but also the most visible and hidden paths between the points of interest. The point was to visualise different routes depending on the intentions of the subject. This illustrates the need to understand human spatial behaviour patterns as two-dimensional probability surfaces rather than points or lines, where there is room for many different possible courses of actions given the same circumstances (Lock & Pouncett 2014).

There has been criticism towards understanding human behaviour through mapped environmental factors within the subject of archaeology. Perhaps the most prominent critic is Mark Gillings, who explained in an article that there is a disconnect between the most avid proponents of GIS technology and the academic researchers who wish to develop explanatory models based on subjective human experience. In the article, the author emphasises the relationship between the features that we are studying and the variables that may explain the presence of said features. This relationship is described as "affordance", which means the ability of the variable to afford the human subject to act in a way, which explains the presence of the studied feature (Gillings 2012). The degree of affordance is bestowed upon the environment by the subject, which makes the importance of environmental variables highly dependent on the subject who perceives it.

Both of these examples show that the relationship between environmental conditions and human decisions is very complex. If the goal is to explain the occurrence of these relationships, it might be impossible due to all the potential factors that might influence decisions or affect the experience that the subject has in the environment. If instead, the goal is to predict occurrences of features, given certain conditions, then explaining the entirety of the complex relationship between features and environmental variables is not the focus. The focus is rather to explore the potential of co-occurrence of features and conditions as an explanatory model to predict one component given the other.

As such, it is important to be specific when discussing the role of archaeological predictive models, they are not frameworks for explaining all the nuances of human spatial experience and behaviour. They are instead tools for locating remains from ancient human activity, which as an effect might aid in the endeavour of further developing our understanding of spatial behaviour and the experiences which might contribute to the observed pattern.

2.3. Tobler's first law of geography

"Everything is related to everything else, but near things are more related than distant things" (Tobler 1970)

This sentence was originally written in the journal "economic geography" in the year 1970, where the geographer Waldo Tobler described population growth simulation for Detroit between the years

1910 to 2000. Tobler recognised that the values of the individual grid cells representing population numbers, which overlay the map of the Detroit area were influenced both by the values of the adjacent cells and the value of the same cells in the previous decade (Waters 2017, pp. 1-2).

The recognition that similarities between features and areas have both spatial and temporal dimensions carries over very well to an archaeological context. The connection between adjacent sites, whether they are contemporary in time or are close in spatial proximity is well understood within landscape archaeology, where researchers refer to this phenomenon as “regions”. Within these regions, it is expected that cultural similarities will be manifested in the archaeological material (Allen & Stanton. 1990, P. 74).

Something that may seem obvious at first but is very relevant when discussing natural features as environmental variables to be used in a spatial analysis, is the tendency of features to have similar characteristics if they are close in proximity. If we can observe a clustered phenomenon when we are studying site distribution in a geographical area, there could be a good argument that there are geographical features which influence the locations of the observed sites, creating the clusters. It is possible to measure the degree of clustering within a GIS environment, i.e., Spatial autocorrelation. I will discuss this further in the Methodology chapter

2.4. Inductive and deductive approaches to predictive modelling:

The methods used when creating a predictive model are one of the essential points of consideration in order to create a high performing tool for assessing the archaeological potential of a given area. The amount of surveying an area has had or how many studies have been carried out beforehand are some of the factors that may determine which kind of model creation method is more appropriate.

There are two main approaches to the creation of predictive models (Verhagen 2018, p.1):

An inductive model, which is based on observed data gathered from a sample of known sites. The environmental parameters that will be researched are determined by the nature of the environmental characteristics of the area. With inductive models, the accessible data determines what understanding we can gain from the model.

A deductive model, in which a hypothesis about the study area is formed. It is based on previously observed and documented information about human behavioural patterns. The hypothesis is formed with assumptions of expected human locational behaviour, which can then be tested through applying statistical tests with a sample. With deductive models, our theoretical understanding about human spatial behaviour determines the quantity and quality of data we can gather from using our predictive model.

When considering which approach to take, there are two main factors to consider:

- Is there a well-developed theoretical framework available for settlement location choice? This is necessary for creating a theory-driven deductive model based on a hypothesis (Verhagen et al 2014, p. 380-381).

- Is there enough data to rely on? When creating an inductive data-driven model, if there isn't a large quantity of data in the sample, the resulting model might not be reliable for predicting sites (Drennan 2009. p. 80).

For identifying the environmental variables with the most predictive power, both approaches could be used. It is important to note that in areas where no surveying has been done, we are reliant on our understanding of prehistoric human locational behaviour, while this isn't the case in well-surveyed areas where we have enough data gathered to make fair assessments on the predictive power of the variables.

The two approaches are not mutually exclusive, and it is important to recognise their compatibility with each other in situations where we are faced with problems such as a lack of data or expected bias. One such example is the distribution of sites in Jutland, where the researcher Ejstrud demonstrates the correlation between site presence and the travel time to the nearest museum (Van Leusen 2005, p.12).
This example shows that we can't solely depend on inductive reasoning when considering which variables may be relevant for predicting the presence of prehistoric sites, as there may exist a bias which affects the physical evidence, which in turn could impair our understanding of the past.

For example, the study conducted in this paper will be inductive in nature, in the sense that the parameters are tested by their ability to predict the presence of the data, which is studied, but also deductive by filtering them out based on how they have been performing in similar environments previously. In the following chapter, the process of assessing the settlement distribution and parameter performance will be presented.

3.0. Methodology

In this chapter I will go through the methodological steps which are necessary to evaluate the predictive potential of the environmental parameters which will be subjected to the analysis, as well as the techniques that are used to complete this task.
The first part goes over the measurement of settlement distribution and whether a clustered pattern can be observed which would indicate that one or more external factors are influencing settlement distribution.

The second part covers the topic of parameter construction, which entails the process of constructing spatial surfaces based on the environmental variables.

In the third and fourth parts I will go over how the estimated performance of the constructed parameters are measured and critically assessed.

To illustrate this process, I will present it in the shape of a flowchart (figure 1), where the process is either disrupted or continues depending on the ability to fulfil a stated criterion. The further down in the flowchart that a variable is able to climb, the more acceptable it is to use in a predictive model, hypothetically. If a variable can pass this process completely, then it is deemed to be suitable to include in a predictive model over the area which is studied.

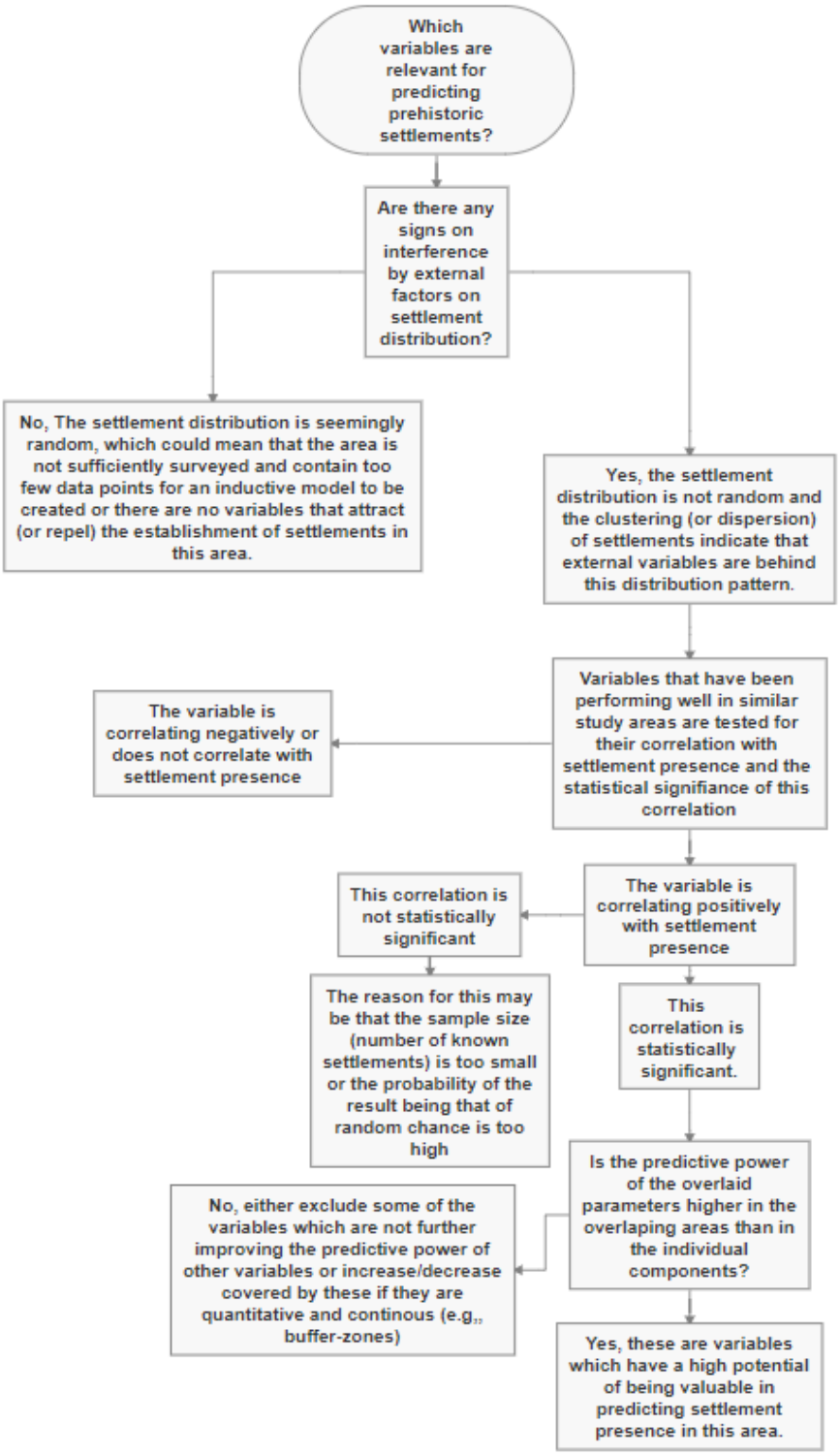


Figure 1: *The flowchart displays the series of methods that will be used to answer the research questions. If a result, which does not support the alternative hypothesis of a statistically significant positive correlation between a variable and settlement presence is displayed, then it is excluded from the analysis, otherwise the test continues. The chart is created by the author of this thesis.*

3.1. Are the sites clustered? Spatial autocorrelation

As I discussed previously in the theory chapter, objects that are close in space tend to share similarities. This connection is called spatial autocorrelation which is defined as "the similarity between observations as a function of the distance between them" (Lucian-Schrader 2013 p. 57). If we have a map over an area with known archaeological sites distributed in it, we want to know whether these are clustered or not, to determine if there may be a parameter which contributes to this pattern of spatial proximity.

To measure the degree of which the settlement distribution is clustered, I have chosen the Global Moran's I test as it is a reliable tool for measuring spatial autocorrelation and has been used in previous studies successfully. One of the earlier implementations of it was in 1990 by Kvamme in the paper "*Spatial Autocorrelation and the Classic Maya Collapse revisited: Refined techniques and new conclusions"*. In this paper, the author used said method to reassess the result from a previous study in which the settlement distribution during the classic Maya collapse was evaluated, with the result of which according to Kvamme "Underscores the need for a careful and thoughtful approach to the analysis of data and the importance of linking appropriate methods with the problem at hand*"* (Kvamme 1990, p. 7).

There are also more recent examples of implementing this method. One example is "*Archaeological Sites in Small Towns—A Sustainability Assessment of Northumberland County"* written by Eric Vaz in 2020. In this paper, he identified settlement hotspots through the implementation of Global Moran's I test on archaeological material (Vaz 2020).

The Global Moran's I test works in the following way:
Within a software environment (for example a GIS), the features whose distribution we wish to study and the surface which contains the features are treated as a single unit in order to calculate the distribution of the surface values. This is easily done with for example elevation data or other raster datasets, where the surface and the data we are studying already are a single unit. I will go through how to do this with vector data later in this chapter.

The result of the analysis is a comparison to an assumed null-hypothesis which states that the distribution of features is random. We therefore get three different values as outputs (URL: https://pro.arcgis.com/en/pro-app/2.8/tool-reference/spatial-statistics/spatial-autocorrelation.htm):

- A z-score, which indicates how far the result is deviating from the expected null-hypothesis of random distribution. A positive score indicates that there is a spatial clustering to some degree, while a negative result indicates that areas with similar values (site presence or site absence) repel each other. A negative result doesn't mean that there aren't any environmental parameters affecting the distribution, but rather that they may deviate from the principle of spatial autocorrelation.

- A p-value, which states whether the observed pattern (or lack thereof) is likely to be a result of chance. A p-value of below 0.05 is generally accepted as being a threshold for statistical significance, which tells us that there is a 5% chance of the null-hypothesis of being true, which we in that case can reject.

- Moran's value, which is a correlation coefficient between the values -1 to 1. If the resulting coefficient is -1, then a perfectly dispersed pattern is observed (Imagine the white or black tiles of a chess board). If the coefficient value is 0, then there is no pattern, and the distribution of features is perfectly random. If the value is 1, all features are observed to be parts of exclusive groups near each other and perfectly separated from other groups (Like the colour distribution of a domino brick).

  All these factors will tell us whether we can say that we can argue for a pattern of spatial distribution, which could be an indication of the influence of an environmental parameter, assuming spatial autocorrelation within the parameter in question being present.

The convenience of using a GIS for this task is that it enables us to efficiently perform and visualise the global Moran's I test to establish what pattern we can find within the study area, what significance it has and what it means for the environmental variables.

First a fine grid of 500 by 500 metres is created to cover the entire study area, each grid unit is given a value based on whether it contains a site or not. A Global Moran's I test is then applied, to test if there is a clustered pattern. Based on Tobler's first law of geography, assuming a similarity in environmental features in areas close to each other, settlements close to each other might have been placed there with that environmental feature in mind. The result of the test gives us a clue of the importance of geographical factors on site presence, as a random distribution (null hypothesis) would be expected over the area if there were no correlations between the geographical features of the area and site presence.

To summarise: Referring to what I mentioned previously in chapter 2.3, if we make the assumption that there is a higher likelihood of the areas spatially closest to any given point share many similarities to that of the area where said point is located than being significantly dissimilar to it, then it is also a higher likelihood that one or more environmental variables can explain the settlement distribution pattern if this is observed to be clustered. If the distribution is shown to be clustered and not the result of random chance, the next step will be to test the performance of the parameters we expect to predict settlement locations. Before that is possible, the parameters have to be constructed. This process will be presented in the following chapter.

3.2. Parameter construction

The creation and testing of the parameters will be performed in ArcGIS Pro 2.7.0, where both analysis and visualisation is possible. It is important to describe the process of how the parameters are created from the environmental variables and explain the different methods to make this kind of study valid and reproducible in the future.

In the cases where the environmental variables are continuous, i.e., defined by different levels of distance from certain features, it is necessary to divide the levels into well-defined intervals. The appropriate division is dependent on several factors, such as spatial resolution (the size of the study area), the type of variable studied and the observed settlement distribution (clustered or not and to what degree). The intervals defined in this paper, will be based on previous research in the cases where this matter has been discussed and evaluated. Buffer zones are created around or from the

features which are the objects of interest, whose dimensions are defined by the given intervals. The number of settlements (defined as points with their positions set as coordinate values), by joining the settlement features with the constructed buffer zones, it's possible to observe the number of settlement features intersecting with the area of the different buffer zones. By knowing the spatial extent of the different buffer zones and the number of settlement features intersecting with these, we have all the fundamental knowledge we need in order to proceed with the analysis.

3.3. Parameter performance assessment

When we are testing the ability of predicting site presence by the individual parameters, the necessary step to take after that is to test how well they perform in conjunction with each other. This will constitute the zones where the parameters overlap, which hypothetically would perform better than areas covered by a single variable. I have chosen the overlay method for this purpose for three main reasons:

- Simplicity: The method is easy to explain and implement within a GIS environment
- Performance: Overlay analysis, especially weighted overlay as a method, is performing well in several different predictive modelling studies relating to archaeological material, such as in Ejstrud (2003) and Nsanziyera (2018).
- Availability: It is possible to perform this method in many different GIS softwares such as QGIS and ESRI's ArcGIS, thus making it replicable for many people who wish to adopt the same method for their study area.

The overlay method relies on two things:

- Data, which is geographically referenced
- A hierarchy of importance assigned to the variables (if the overlay is weighted)

The logic behind the method is the following: If we have two different variables whose presence affects human survival in a positive manner, then it would be preferable to settle within either of these two as opposed to settling in an area absent of either of them, but it would be even more preferable to settle where both of them are present at the same time.

The performance value is expressed in Kvamme's gain, which tells us the degree of settlement quantity overrepresentation in a given area in relation to its size. The hypothesis is that areas where environmental variables intersect have a larger overrepresentation than its individual components, assuming equal importance between all variables.

Assuming equal importance for all variables is misleading, therefore a weighting is usually performed, where either expert opinion is involved, or we rely on empirical data as a basis for testing the level of statistical significance. The goal of this study is to test and discuss parameter performance, not to create a predictive model, which is the reason why a weighting will not be performed. The performance of the environmental variables at different spatial levels are tested, which might alleviate future research in assigning weights to these variables.

3.4. Parameter significance

When we have established which parameters we wish to include in the construction of our predictive model, there is a convenient method to use for determining whether the correlation between the variable and settlement presence is strong enough to not be considered a result of random chance. This is called the Kolmogorov-Smirnov single sample statistic (KS).

Using this method on archaeological material has been successful previously, for example in the paper "GIS and Remote-Sensing Application in Archaeological Site Mapping in the Awsard Area (Morocco)". In this work, the authors tested both the statistical correlation between site presence and many different environmental parameters along with the significance of these correlations. They did this through the following procedure presented below (Nsanziyera 2018, pp. 9-11):

First, we make subdivisions within the individual parameters, Secondly, an assumption is made that the percentage of sites covered by the parameter directly corresponds to the percentage of land a parameter covers, this is our null-hypothesis which we will try to reject, with the alternative hypothesis being that there is a significant deviation from this null-hypothesis. In KS terms, the null hypothesis is our expected cumulative frequency. Thirdly, an empirical observation is done to establish how many sites are covered by the respective subdivisions, which is called the observed cumulative frequency.

The difference between the expected and the observed frequencies is expressed as a d-value. The higher the d-value is above a defined threshold (D), the more the observed frequency differs from the expected frequency and the null-hypothesis could potentially therefore be rejected.
The threshold value depends on the size of the sample. The larger sample size, the less convincing the correlation between the dependent variable (settlement presence) and the independent variables (the studied parameters) need to be in order to be deemed as not a result of random chance.
To determine the threshold D-value we use the following formula (Lee 2005):

$$"D = 1.358 / \sqrt{N}"$$

Where 1.358 is a constant and N is the number of known sites in the study area. The constant depends on what margin of error we wish to apply, 1.358 corresponds to a p-value of 0.05, which is a standard marker for there being a 5% or less chance of the result being one of random chance (Massey 1951, p. 4).

The formula for the critical value D is derived from the assumption that the expected behaviour of a statistic is a function of the sample size, which is called "large-sample theory". This theory states that the larger sample we have, the more accurate the values representing the observed correlation between the dependent variable and independent variables are (Lockhart 2020). This means that even if we don't have access to the knowledge of all settlements which has existed in a given area, the correlation between the presence of settlements that we are aware of and the parameters we are studying is expected to be similar to the actual value if we knew the locations of all settlements that have existed in that area, given a large enough sample size.

If the D statistic is lesser than any of the d values of the individual subdivisions of the parameters, then we can reject the null-hypothesis and say that the settlement distribution has something to do

with the parameter in question. If none of the subdivisions has a d-value greater than the threshold D-value, then we cannot disprove the null-hypothesis.
In essence, there needs to be either a very strong observed correlation between the dependent and independent variables or a moderate correlation with a large sample size to make the argument that the correlation is statistically significant.

4. Materials and methods

In this chapter, the material, which is studied, along with the methods applied to these in order to perform the analysis is presented. First, the study area is described and put into context with the surrounding areas, with a brief overview of the area's prehistoric cultural history. Secondly, the sources from which the data is gathered are presented, finally the variables that are chosen based on previous research are presented with descriptions of the processing methods.

4.1. The study area

The study area is in north-western Scania and covers a land area of 1 063,46 km^2 (excluding the area covered by major lakes). To the west, the area is bordered by the Danish straits. To the north, there is the border between the counties of Scania and Halland. The southern and eastern borders are arbitrarily drawn by the author to get a study area of an appropriate size containing an appropriate amount of data points (settlements).

The area is chosen for two reasons:

- The number of already identified settlements within the area gives an adequate sample size of 87 data points. The density of settlement sites was regarded as sufficient for limiting the area size, while being able to work with a sample close to 100 data points.
- The relative topographical and environmental diversity of the study area compared to other considered parts of Scania provides an interesting testing ground for the components of a predictive model.

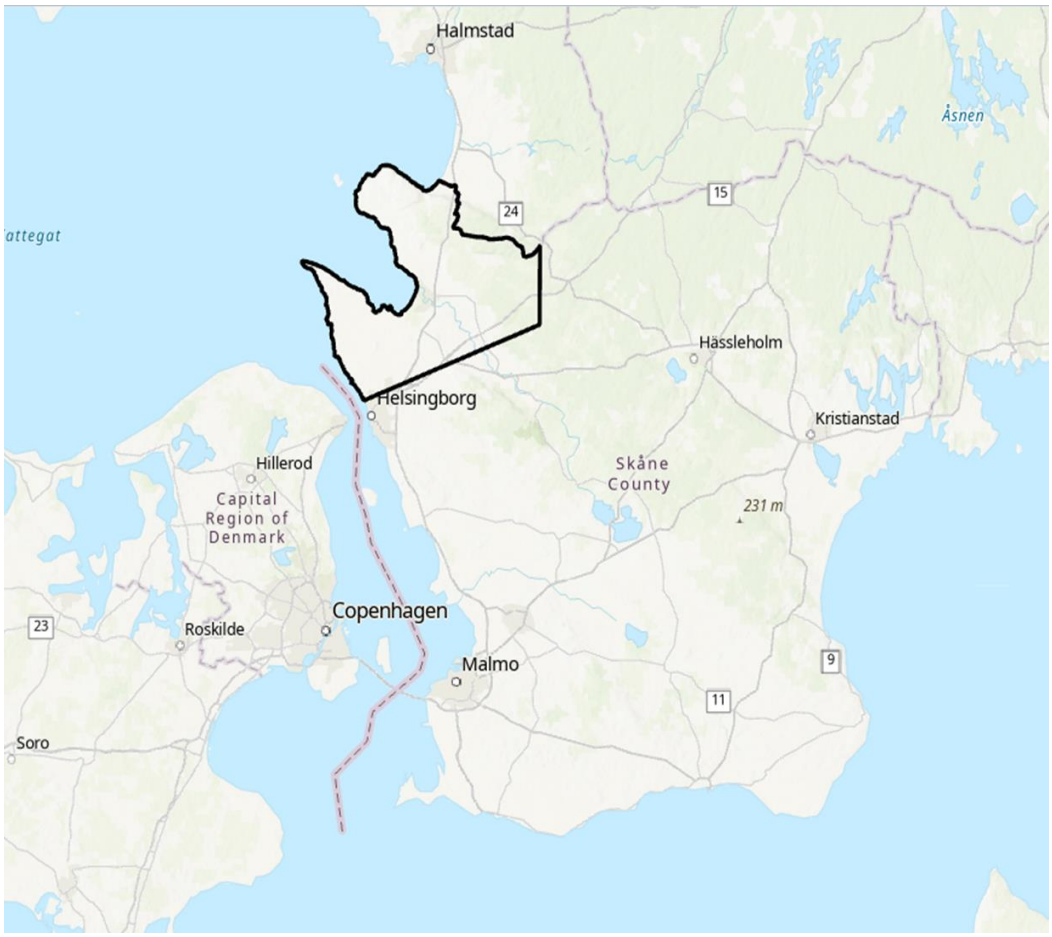


Figure 2: The map displays the location of the study area within the county of Scania in a black outline.

Below follows an overview of the different parts of the chosen study area (Blomberg & Helgesson 1996 p.134-140):

Between the mountainous peninsula of Kullaberg in the north-western part of the study area, the hills of Söderåsen in the eastern part and Hallandsåsen in the northern part, the area is characterised by coastal plains and river basins (Berglund & Rapp 1988 p.19).

In the middle of the area the plains of Ängelholm are situated, where the river Rönne å flows through. This alluvial plain hosted settlements along Rönne å already in prehistoric times, a tendency which seems to have had a continuity into later ages, judging by the many villages and churches that were located along the river during the Middle Ages (Blomberg & Helgesson 1996).

Being exposed to extensive agricultural activity over the many years of human habitation, the many wetlands and small rivers that were present before industrial agriculture are now mostly gone, giving place to arable land.
This plain is somewhat of a geological anomaly, due to the presence of most of the postglacial sand in Scania, which is otherwise not a common soil type in this part of Sweden (Helgesson 2002 p.7).

One of the major post glacial ridges of southern Sweden runs through the northern part of the study area, called Hallandsåsen. This region was mostly sparsely populated during prehistoric times apart from the western edges of the ridge near the coast. The topographical variation in this area is the most extreme in the entire study area, exceeding 150 metres.

In the western part of the study area the region is characterised by fertile soil in the central and southern parts, along with topographically elevated positions in the northern parts. The Bjäre peninsula is in the north-western part of the study area, which with its great overview over both the coast along with the fertile plains to the south, contains a rich amount of prehistoric material remains. In the south-eastern part of the study area, the north-western parts of the glacial ridge of Söderåsen is located.

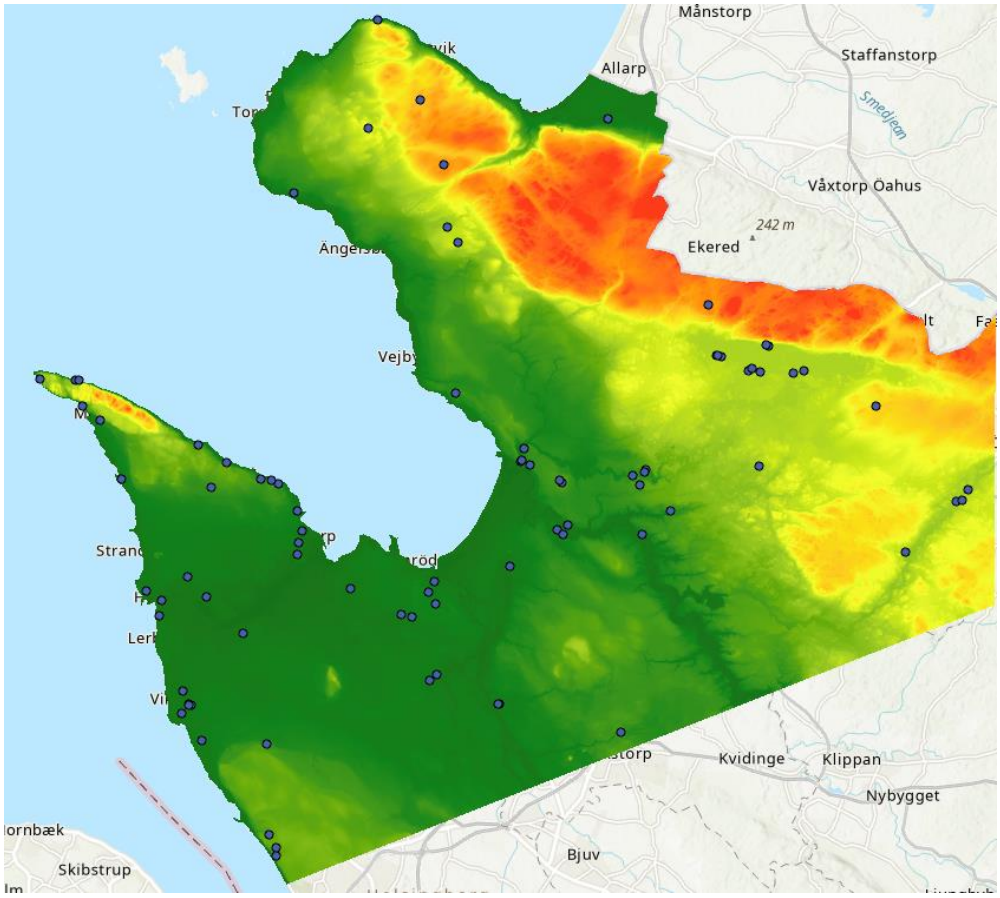


Figure 3: Topographical overview over the study area, with elevation ranging from 3 metres below sea level (dark green) to 211 metres above sea level (dark orange). Observed settlements as registered by Riksantikvarieämbetet are represented by blue dots. The hilly Bjäre-peninsula is located in the western part, while the majority of the northern study area is covered by the hills of Hallandsåsen to the north. The northwestern part of Söderåsen stretches into the southeastern part of the study area. Sources: Lantmäteriet, Riksantikvarieämbetet

4.2. Riksantikvarieämbetet

The settlement data that will be used for testing the predictive power of the environmental parameters is prehistoric settlement data from the Swedish national heritage board, "Riksantikvarieämbetet". In recent years the organisation has digitised registered archaeological data along with the geographical position of the finds, making it accessible for downloading and processing in a geodata-format.
The category of sites which will be used are sites which have been explicitly defined as settlements. Due to the difficulty in dating the settlements, with no written record of this matter in the material for a distinction to be possible, there will be a low temporal resolution with all time periods represented in a single sample.

4.3. Lantmäteriet/Land Survey Institution of Sweden

The land survey institution of Sweden has a database containing geographically referenced information ranging from satellite photographs to features such as rivers and land use. This will be the source from which the environmental parameters are created. All of the collecting and processing is done by the Land survey institution, who are the owners of all data which will be used to create the parameters in this study. The Swedish university of agricultural sciences is the controller of the data, from whom it is downloaded. Some of the data is the result of instrument measurements, such as aerial laser scanners, while others are drawn features in a digital environment, such as rivers and lakes. Thus, the accuracy can vary depending on how the data has been collected or processed.

4.4. The environmental parameters

As has been demonstrated through the previous studies presented in this paper, the most common method of choosing which variables to test for predictive power in any given area is to examine what has been done previously in the study area regarding this very topic. This also includes variables which has worked in study areas of similar characteristics, whether it's geographical proximity or physical similarity of the landscape. As I have mentioned earlier, to my knowledge, an archaeological predictive model over Scania hasn't been done, which leaves the second option the only one available.

The following variables have been included in the 6 studies which has been done in regions close to Scania:

Soil type: 6
Distance to water bodies: 4
Distance to coast: 1
Presence of wetlands: 1
Elevation: 2
Slope: 3
Aspect: 2
Geomorphological type: 1
Density of sites/proximity to other settlements: 2
Historical maps of land use: 1

Three of these case studies are Swedish, while one is Dutch, one is German, and one is Danish. It's interesting to note that very few of the variables that have been included in said studies have differed significantly from the others. The reason for this could be that all authors have based their choices on early academic works by other authors such as Kvamme. It could also mean that they have identified the same variables as the most appropriate ones for their particular study areas independently from each other, due to the common geographical characteristics that their study areas have.

The successful use of these variables as basis for their predictive models leads to the conclusion that they would be appropriate to use as predictive parameters for a model created for a Scanian setting. The implementation and testing of this will be the objective for the rest of this paper.

The following variables have been selected from the sample mentioned above:

4.4.1. Soil type

The only type of variable that was present in all 6 case studies seems to be the most common variable in archeological predictive modelling in general. The only exception is when the studied material is from cultures which relied heavily on hunting and gathering or if the terrain is unable to host agriculture, such as in the Piñon project in Colorado (Kvamme 1992).

The temporal resolution of this study is low, which means that the material is from time periods where agriculture wasn't present along with material which was. With that said, soil type as a predictive variable is a reliable choice for a building block when constructing a predictive model over Scania, due to the long history of agriculture, spanning many time-periods.

The types of soil that will be tested are rock, glaciofluvial soil, silt clay, moraine, moraine clay, postglacial sand and peat. The spatial resolution is 1: 1 000 000. This is a rather low resolution, but it was the only one available that covers this region of Sweden to my knowledge. The problem with low resolution will be a matter of further discussion later.

4.4.2. Distance to water sources

As has been observed in the previously mentioned studies, a close distance to sources of freshwater is important for predicting the presence of prehistoric sites. The reasons for this could be that there is a constant demand for it both as drinking water and for agricultural purposes. This is a variable that has been significant regardless of time-period, which makes it especially appropriate for a study with low temporal resolution. The distances that will be tested are based on what previous studies have shown are successful, but there will be a test for the optimal distance based on the empirical evidence at hand.

Due to the changing agricultural landscape of the region, it can be advantageous to use historical maps as complementary information for mapping the waterways, whose appearance and extension has changed significantly over the years. Between the years of 1805 and 1914, the total amount of arable land in Scania increased from 13 to 54%, this increase came as a result of the drainage of waterways and lakes during this time (Blomberg & Helgesson 1996. p.52).

In this study, a historical map from 1911, created by the General staff of Sweden was used to identify the major waterways which have been drained during the last century. The map is georeferenced based on common features with the current study area, primarily using peninsulas and distinctive coastline features. It would have been preferable to use an older map, due to the extensive drainage prior to its creation, but it should fulfil its purpose adequately in this study. The map covers most of the study area and has many clearly distinguishable rivers, which cannot be seen in modern maps.

The appearance of many of the currently existing waterways are different from those seen in the map as well, which was considered when recreating the waterways in the GIS. In the cases where rivers were not distinguishable from roads (due to sharing the same colour), they were not included in the construction of the dataset.

After the lakes and rivers were manually drawn and the variable was created, the dataset was ready to be subjected to the analysis. 5 different buffer zones were then created around all water bodies: 100, 200, 300, 400 and 500 metres from the variable were the distances selected.

A 500-metre distance has been considered sufficient according to previous experimentation in a study area like the one in this study (Ducke 2004, p.2), which will be considered and be tested to verify if this indeed is the case in this study area.

4.4.3. Distance to coast

Distance to coast is an important variable from a transportation viewpoint. The sea connected areas and communities rather than separated them, thus it is an appropriate variable to take into consideration. This variable hasn't been explicitly mentioned in the case studies often, but rather has been used interchangeably with distance to water bodies, which is understandable due to the ability to transport through rivers and waterways to the coasts.

As mentioned previously, the settlements that are studied in this paper are from a wide variety of time periods, during which the coastline has been constantly changing. This has two main causes: local crustal depression or uplift and global eustatic sea level lowering or rise. The complex relationship between these two variables makes it difficult to give an exact estimate for any single location in the Fennoscandian geological region at a given time (Påsse & Daniels 2015 p.6).

The Baltic Sea has gone through several different periods of damming and flooding between 13,500 and 8900 years before present which made the region oscillate between being a lake and a sea connected to the Atlantic Ocean. This was until an outlet was formed where the modern Danish Straits are, and the modern Baltic Sea started to take shape (Påsse & Daniels 2015 p.13).

4.4.4. Slope and aspect

Despite Scania being characterised by a rather flat landscape, slope could be an important factor if the resolution of the elevation data is high enough. In this study, a DEM with a surface containing squares with the dimensions of two-by-two metres, representing a generalised elevation value for the entire area covered by each square is used.

The slope is then calculated within the GIS software by estimating the necessary slope between each square by interpolation, for explaining the observed difference in elevation values. The accuracy of the result is highly dependent on the spatial resolution of the dataset. Although high resolution is important to yield accurate results, the larger the study area and the higher the resolution, the greater the required hardware processing power is (Chapman 2006 p. 77). This will be a topic of further examination in the discussion chapter.

After creating the slope parameter, the aspect of each surface component can be determined, which refers to the direction of which the slope is facing. A common hypothesis is that it would be preferable to settle in a location where there is a maximal amount of sunlight exposure for agricultural efficiency. One example of a study where this parameter was taken into account is in north-eastern Romania, where the researchers studied Neolithic settlement locations with respect to, among other parameters, the slope aspect of the landscape (Nicu et al 2019).

Southwest facing slopes were considered to have a high value of probability of containing sites, with northeast facing sites were considered to have a low probability. This is called the "heat load index", with the areas with most exposure to sunlight being the areas which have the most heat load, being the most preferable regions for agriculture (Nicu et al 2019 p.4).

4.4.5. Proximity to known settlements

This variable is different from the others, it isn't grounded in the hypothesis of spatial autocorrelation, but rather a tendency for human settlement patterns to be continuous in a single area over a long duration of time. While proximity to already known settlements might be a self-fulfilling prophecy, due to the possibility of their discovery being a result of surveying areas around settlements known before them, it may be necessary to take regional-settlement continuity into consideration when attempting to predict where unknown sites might be.

This parameter will be constructed in the following way: A random selection using the "random" module in the programming language Python is done, where a series of 5 random numbers between 1 and 87 (the number of settlement features) are selected. The corresponding settlement index with a randomly generated number is selected and circular buffer zones are created around them set at different distance intervals. As described in the "Parameter construction" chapter, these zones are then joined with the settlement features, from which point we can analyse the relationship between the spatial extent of the buffer zones and the number of settlements contained within them.

This procedure is done five times to create a set of 25 randomly selected settlement features, whose performance to predict the presence of other settlements will be tested, in this test, the settlement from where the buffer zone is created, will not be counted. The Kvamme's gain values and d-statistic will be calculated from the average performance between all five iterations and be the basis for the result and performance evaluation of this parameter.

5. Analysis

In this chapter the methods mentioned previously, are applied to the material. Starting with measuring the degree of clustering, then moving into testing the created parameters. Finally, the result is presented.

The analysis is performed with the following hypotheses:

- The settlement distribution shows a clustered pattern
- The parameters tested in this chapter contains an overrepresented number of settlements compared to their areal extent
- Areas where there is an overlap between one or several parameters contain a larger overrepresentation of settlements than any of the overlapping parameters individually

5.1. The degree of settlement clustering

Before analysing the significance of the parameters, spatial autocorrelation is measured in order to determine whether there are existing settlement patterns which could imply a correlation with environmental parameters. There are a few disclaimers to be addressed before presenting the results of the Moran's I test applied to the material within the study area:

- The cell size of the grid is arbitrarily chosen as a compromise between maximising the number of cells displaying settlement presence and keeping visual clarity when presenting the result.

- There are several settlements remains within a few of the cells, this is not considered in the analysis, due to the following reasons: The Moran's I test is testing for proximity of similar values, which would entail that a cell containing exactly one settlement is as dissimilar to a cell containing two settlements as a cell containing 0. Because we want to measure the clustering of the settlement distribution, it's more appropriate to have a Boolean system of either presence or absence.

- Proximity in this case is determined by immediate neighbours within 500 metres (a single cell width and height). The higher probability of any given cell on the grid having the same value as its neighbours, the more clustered the settlement pattern is calculated to be, while the opposite is true if there is a low probability of having the same value as its neighbours.

- The more evenly the study area has been surveyed, the more accurate this assessment of spatial distribution is. The analysis assumes that the cells without sites don't contain sites, rather than being unsurveyed.

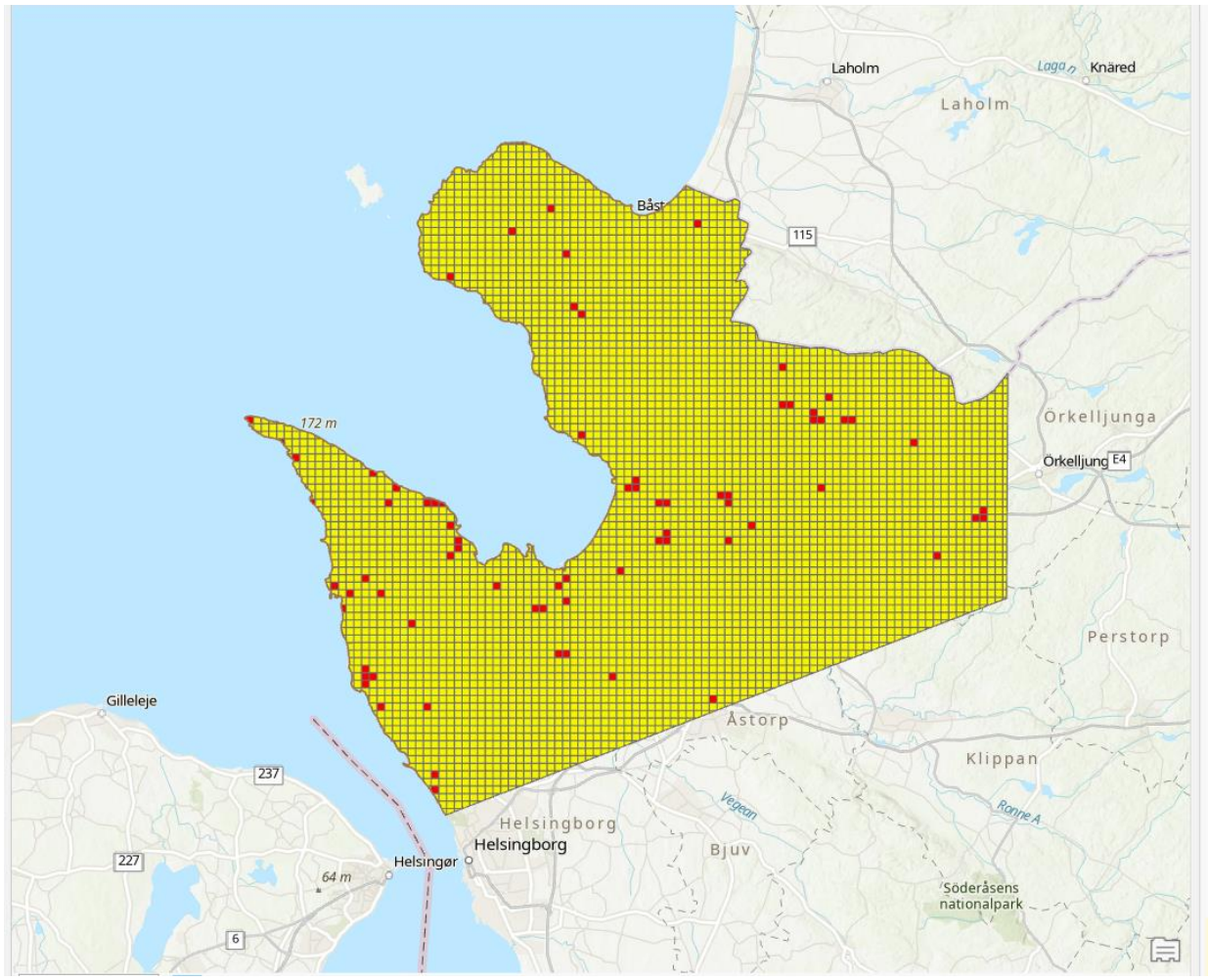
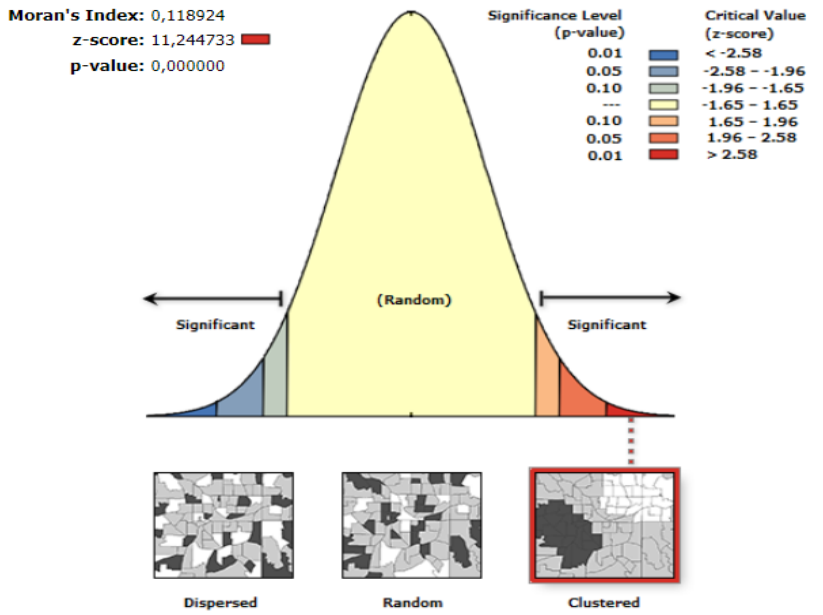


Figure 4: *A grid covering the study area with a cell size of 500x500 metres. The yellow areas are cells where the presence of settlements isn't known. The red areas are cells where there is at least one settlement present.*

Spatial Autocorrelation Report



Given the z-score of 11.244733, there is a less than 1% likelihood that this clustered pattern could be the result of random chance.

Global Moran's I Summary

Moran's Index:	0,118924
Expected Index:	-0,000219
Variance:	0,000112
z-score:	11,244733
p-value:	0,000000

Figure 5: *The result of the Global Moran's I test, including the degree of clustering of the settlements, as well as the statistical significance of the result.*

The result of the Global Moran's I test shows that the settlement distribution is clustered (figure 5), although not by a large margin, given an Index score of approximately 0,12 on a scale from -1(perfectly dispersed), 0 (perfectly random distribution) and 1 (perfectly clustered). The probability of the settlement distribution to be attributed to random chance is presented as 0%, stated by the p-value of 0,000000, which probably should be interpreted as an approximation rather than an expression of exact truth.
The extremely high z-value of 11,2 also points in this direction, i.e., the probability that this distribution is randomly generated is extremely low, less than 1 % chance.

The pattern of settlement distribution which we can observe from the result of the analysis is with a very high likelihood implying that there are one or more dependent variables affecting it. To emphasise the significance of this result, a randomised version of the distribution map is created to compare the results of both the map generated by the empirical data and the random point distribution.

This is done by generating points with random locations distributed within the study area, which are overlaid by a grid with the same size as in the empirical test. The same procedure is then carried out as previously, generating a map over randomly distributed settlement locations in the study area (figure 6). After running the Global Moran's I test on this distribution, the result (figure 7) validates the legitimacy of the result from the empirical test, showing the difference between what a surface containing randomly generated data and the same surface containing the actual data.

The comparison between the results of the Moran's I test applied on the empirical and the random distribution maps shows that the probability of the empirical data to be the result of random chance is extremely low, and thus the null hypothesis of a random distribution with no influence from external dependent variables can be rejected. To further validate the significance of this result, there should be dozens of randomised distribution maps generated, but due to a lack of time and the convincing nature of the comparison between the results that has been presented here, no further validation will proceed at this time.

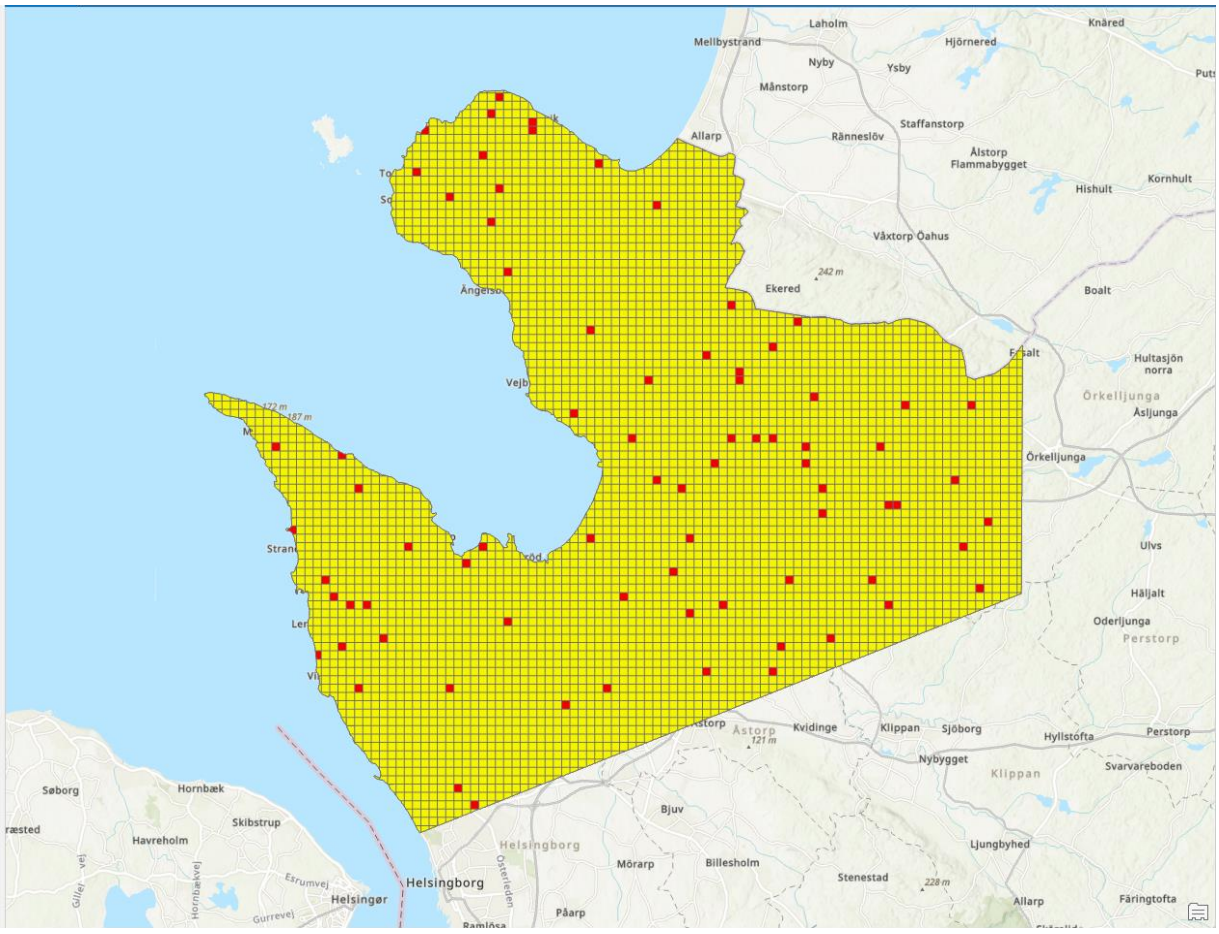


Figure 6: *A grid covering the study area with the cell dimensions of 500x500 metres, where a set of 77 random points have been generated. The red cells contain the randomly generated points, and the yellow cells represent areas without points.*

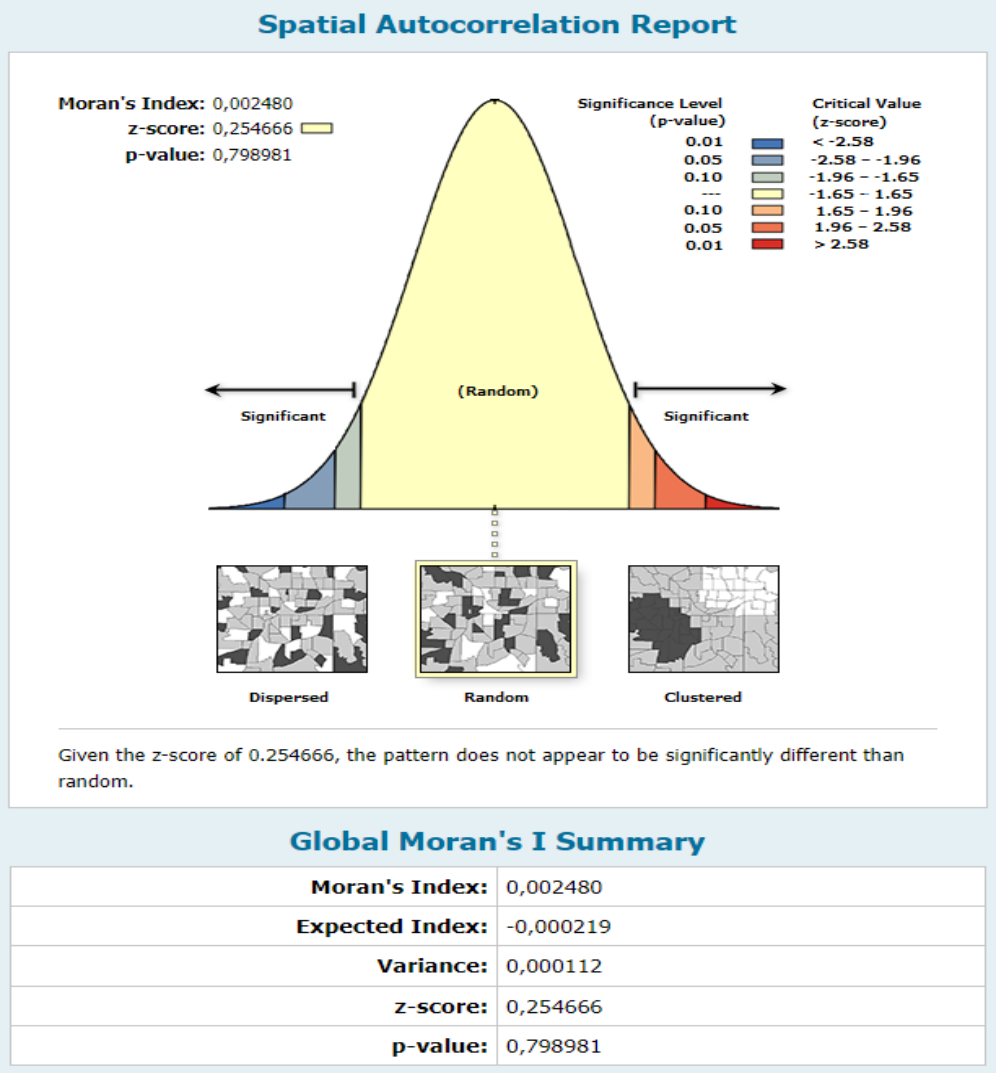


Figure 7: *The result from the Global Moran's I test when applied to the randomly generated dataset. No clustering is observed, and the points are randomly distributed over the area, thus no traces of interference by independent environmental variables on the spatial distribution is implied.*

5.2. Testing the parameters

As mentioned previously in the parameter performance assessment chapter, the variables will be tested by their internal performance to predict site locations among their subdivisions. This is done to assess whether the variable at any point during its continuity exceeds the expected value indicated by the null hypothesis to a degree where statistical significance can be asserted.

To illustrate this, I will give an example:
We have an imaginary continuous variable X with five different subdivisions: X1, X2, X3, X4 and X5.

X1 covers 10% of the study area, X2 covers 15%, X3 covers 20%, X4 covers 25% and X5 covers 30%.
X1 contains 15% of all settlements of the study area, X2 contains 15%, X3 contains 40%, X4 contains 10% and X5 contains 20%.

Our null hypothesis is that all subdivisions ranging from X1 to X5 contain as many sites as the percentage they cover of the entire study area. We want to know if this is false and, in that case, how false it is. This is where the d-value comes in, which states the difference between the empirically observed result and what we would expect by the null hypothesis.

X1: d = 0,15 - 0,10 (0,05)
X2: d = 0,15 - 0,15 (0,05)
X3: d = 0,40 - 0,20 (0,20)
X4: d = 0,10 - 0,25 (- 0,15)
X5: d = 0,20 - 0,30 (- 0,10)

The expected frequency corresponds to the percentage of area the variable covers if the null-hypothesis is correct, and the cumulative frequency corresponds to the actual percentage of settlements covered by the different subdivisions of the variable.

Assuming there are 100 data points (N), the Kolmogorov-Smirnov D-value which would be defined as the threshold for significance is 0,136 (1.358 / √N). As we can see from the example, only the difference between the cumulative and expected frequency in X3 exceeds the assigned threshold value of 0,136, which is enough for us to call the correlation between the whole variable X and settlement presence statistically significant and therefore reject the null hypothesis.

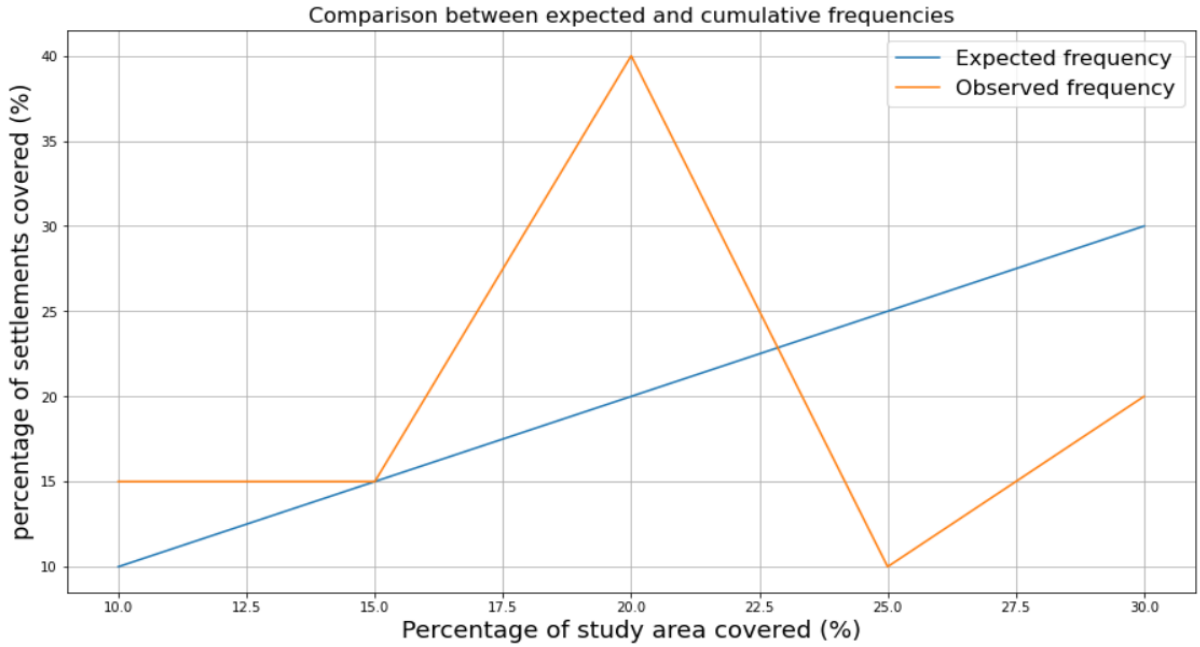


Figure 8: *The chart displays the difference between the expected and cumulative frequencies of variable X. if the vertical distance between both plots at any point exceeds 13,6 percentage points, variable X can be considered as having a significant correlation with settlement presence.*

In this study, the sample size is 87, thus all calculated d-values of the subdivisions of the variables will be compared to the referent D-value of 0,146.

5.2.1. Postglacial sand

The different soil types which are subjected to the analysis are: rock, glaciofluvial soil, clay-silt, moraine, moraine-clay, postglacial sand and peat. The only category which performed above the D statistic threshold of 0,146 is postglacial sand, which scored a result of d = 0,199. The category covers 18,1 % of the study land area (excluding major rivers and lakes), while containing 37,9 % of the settlements, scoring a Kvamme's gain of 0,52. The reason behind the significance of this subdivision is unclear, although the geological layer seems to be a result of withdrawal by an ancient shoreline according to the Geological survey of Sweden (URL: https://www.sgu.se/om-geologi/jord/fran-istid-till-nutid/landhojning-fran-havsbotten-till-lerslatt/postglacial-sand-och-grus/).

It is worth emphasising that the dataset which was used is in the scale of 1: 1 000 000 applied on a study area smaller than the dataset is intended to, which might affect the result. Nevertheless, the category of post-glacial sand-gravel will be included in the final analysis and the potential use of this variable for further creation and validation of predictive models in this area will be discussed in the discussion chapter.

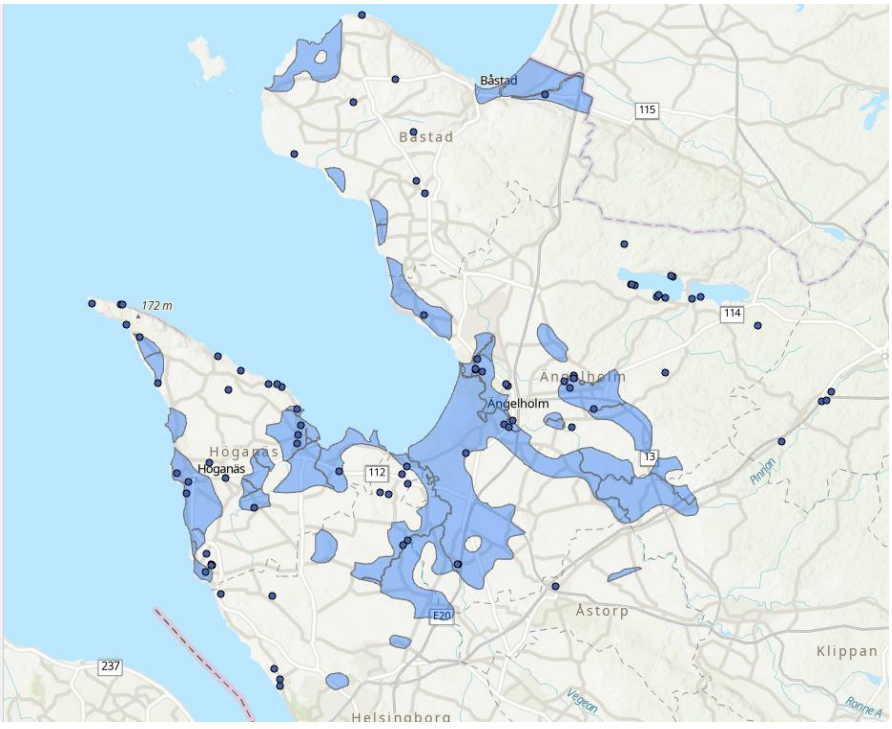


Figure 9: *The extent of postglacial sand in the study area, displayed in transparent blue colour. The location of the settlement locations are represented by points. Sources: Lantmäteriet, Riksantikvarieämbetet*

5.2.2. Distance to the coast

The buffer zones for the distance to coast variable are larger than the other variables tested in this study, due to the low temporal resolution which means that the settlements could be from many different time periods and the coastline could have been at many different levels.

Five different distances were subjected to the analysis: 500, 1000, 1500, 2000 and 2500 metres. All distances were calculated from the modern coastline without considering different conditions during prehistoric times. The implications of this will be part of the discussion chapter.
The distances were mapped by creating multiple buffer zones into the study area from the coastline, then the buffer zones were spatially joined with the locational markers of the settlements, which formed the foundation of the analysis.
Due to the variable being continuous, each distance interval is tested for their d-value and Kvamme's gain.

All distances performed very well in relation to the D statistic of 0,146. The analysis yielded the following d-values:

- 500 metres: The buffer zone covers 5,3 % of the study area and includes 24,1 % of the settlements. Gain = 0,78, d = 0,188
- 1000 metres: The buffer zone covers 10,2 % of the study area and includes 34,5 % of the settlements. Gain = 0,7, d = 0,243
- 1500 metres: The buffer zone covers 15,0 % of the study area and includes 43,7 % of the settlements. Gain = 0,66, d = 0,287
- 2000 metres: The buffer zone covers 19,5 % of the study area and includes 47,1 % of the settlements. Gain = 0,59, d = 0,273
- 2500 metres: The buffer zone covers 23,9 % of the study area and includes 49,4 % of the settlements. Gain = 0,52, d = 0,252

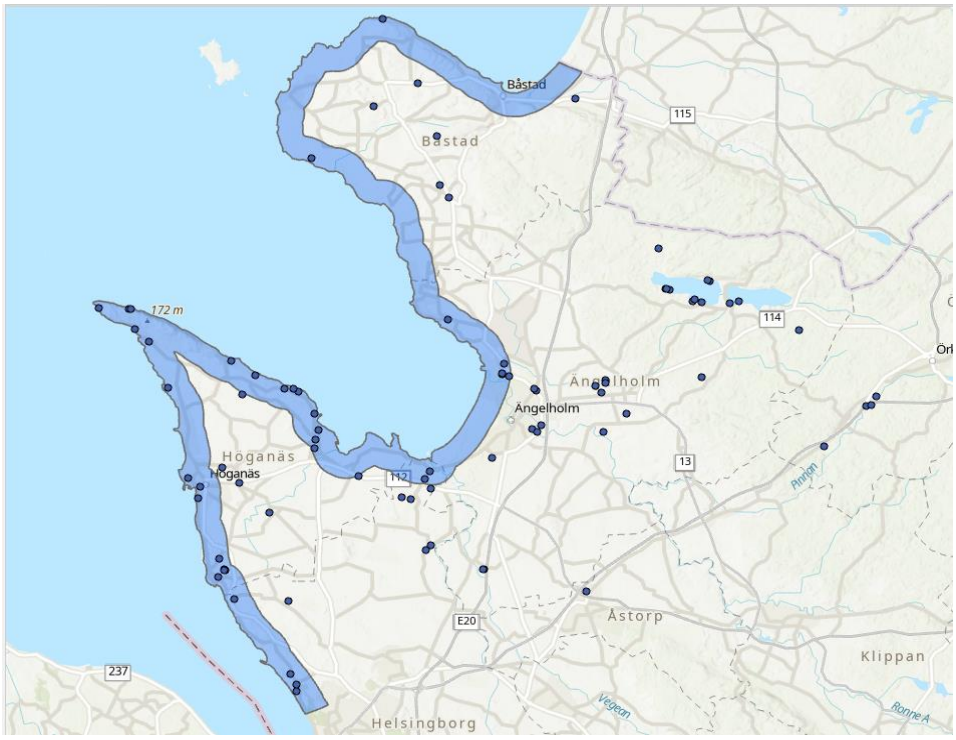


Figure 10: *The extent of the 1500 metre buffer zone from the coast in the study area, displayed in transparent blue colour. The location of the settlements is represented by points. Sources: Lantmäteriet, Riksantikvarieämbetet*

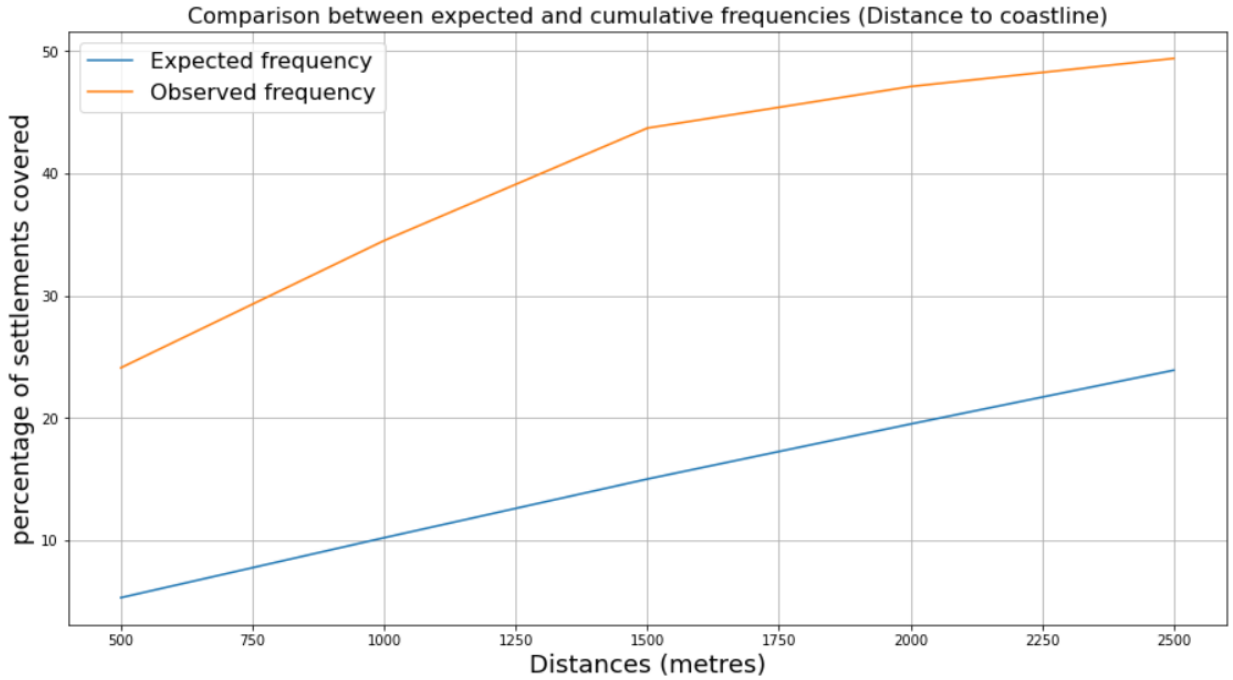


Figure 11: *The chart displays the difference between the number of settlements covered depending on the size of the buffer zones. The blue line represents what we would expect to see according to the null hypothesis of the percentage of area covered directly corresponding to the percentage of settlements contained. The orange line shows the observed values of the percentage of settlements within the set distances.*

All subdivisions of the distance to coast variable performed above the threshold. The distance of 1500 metres is the subdivision which deviated the most from the expected level according to the null hypothesis. While the buffer distance of 500 metres performed better than all the other distances in Kvamme's Gain, it might be advantageous to use a parameter that covers a larger area, a subject which I will discuss further later.

5.2.3. Distance to lakes and major rivers:

Akin to the distance to coast variable, all buffer distances performed well above the D-statistic threshold:

- 100 metres: d = 0,213. The buffer zone covers 6,3% of the study area and includes 27,6 % of the settlements. Gain = 0,77
- 200 metres: d = 0,237. The buffer zone covers 10,8% of the study area and includes 34,5 % of the settlements. Gain = 0,68
- 300 metres: d = 0,240. The buffer zone covers 15,1% of the study area and includes 39,1 % of the settlements. Gain = 0,61
- 400 metres: d = 0,243. The buffer zone covers 19,4% of the study area and includes 43,7% of the settlements. Gain = 0,56
- 500 metres: d = 0,247. The buffer zone covers 23,6% of the study area and includes 48,3% of the settlements. Gain = 0,51

As mentioned in the Material and methods chapter, the 500-metre buffer distance suggested by Ducke (Ducke 2010 p.2) proved sufficient for predicting settlement presence, yielding a Kvamme's gain of above 0,5 and a d-statistic well above the threshold of 0,146.

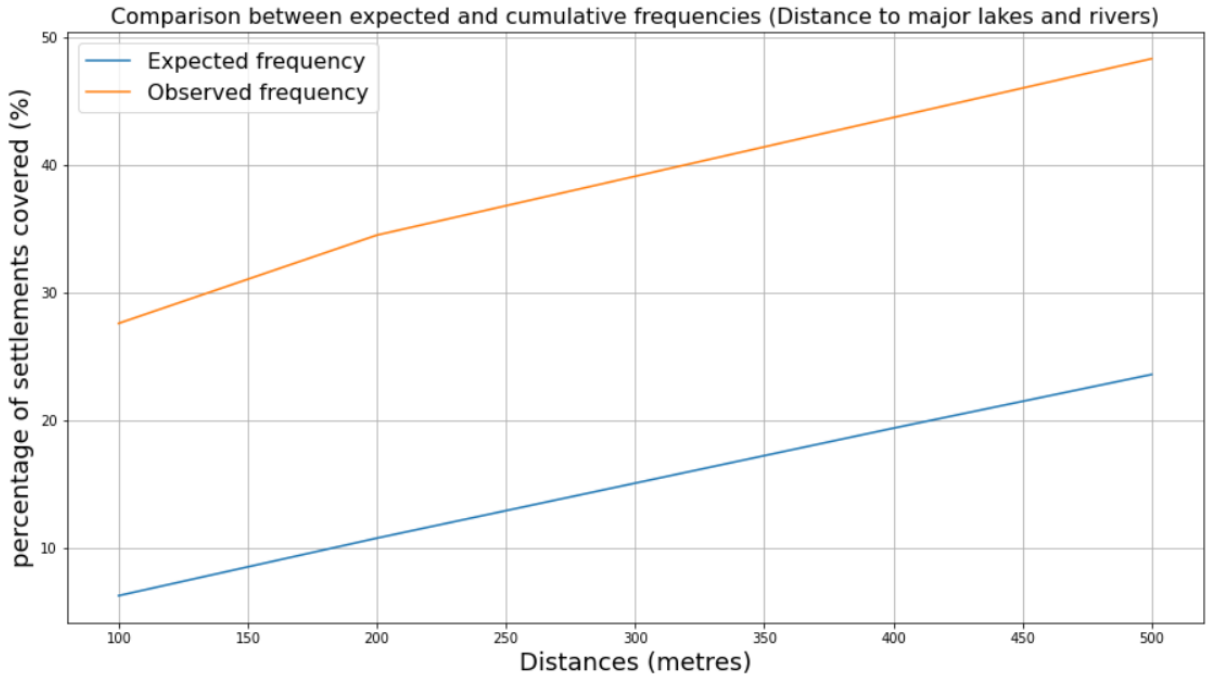


Figure 12: *The chart displays the difference between the number of settlements covered depending on the size of the buffer zones. The blue line represents what we would expect to see according to the null hypothesis of the percentage of area covered directly corresponding to the percentage of settlements contained. The orange line shows the observed values of the percentage of settlements within the set distances.*

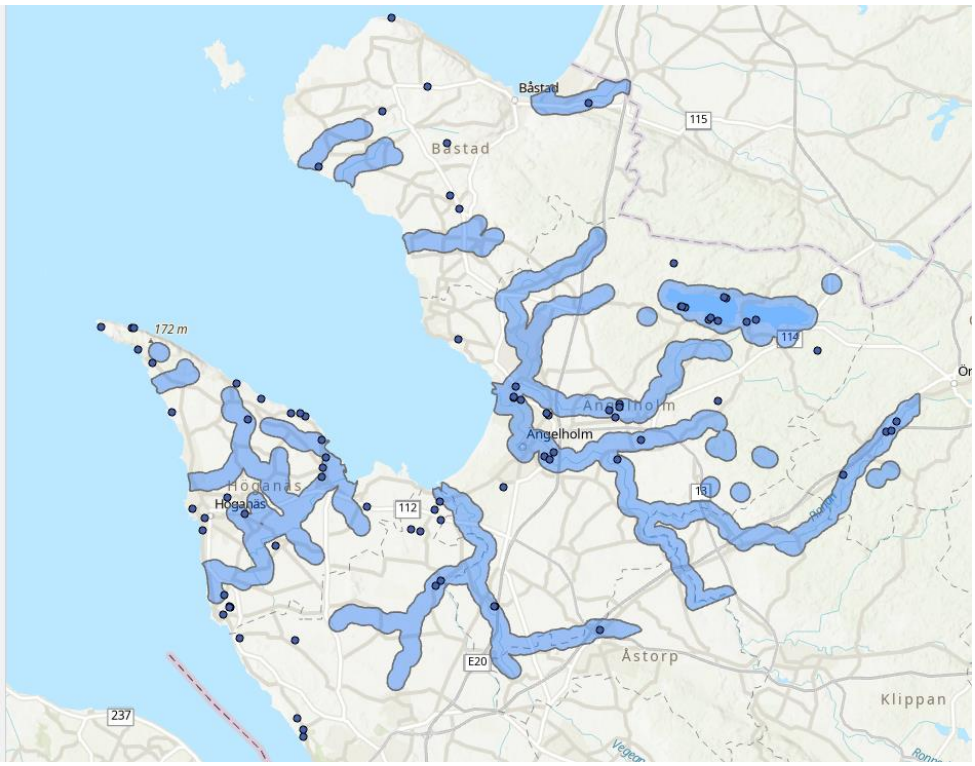


Figure 13: *The extent of the 500-metre buffer zone around the major lakes and rivers in the study area, displayed in transparent blue colour. The location of the settlements are represented by points.*

5.2.4. Distance to settlements

Judging from the high Kvamme's gain values, this parameter performed extraordinarily well, containing a very high percentage of settlements within it at any given distance interval, especially within the 500 to 1000 metre distances, where they display values way over many predictive models. At all set distances, the variable also greatly exceeded the D-statistic threshold of 0,146. However, this parameter has a few major shortcomings, which I will describe below:

It is possible that the settlements who are parts of different clusters have been discovered as a result of surveying these general areas at a higher degree than other areas. This makes the use of this parameter as a tool for predicting site presence a self-fulfilling prophecy, where areas displaying a high degree of settlement clustering being given more extensive surveying, which further enforces the notion that these are richer in site density than other areas.

The small extent of the buffer zones and the small number of settlements contained within these, makes it unreliable for use in conjunction with other variables for excluding

The variable is rather unsuitable to use in an overlay analysis. The reason for this is that we either must randomly select a set of settlements around which we wish to draw the buffer zones or include all of them simultaneously in the overlay. The first option will make the performance of the model unreliable, due to the dependence of a successful random outcome. The second option is more feasible, but if the settlement distribution is too clustered, the result will be skewed due to the buffer zones containing each other's settlements. This would mean that a rather large surface of the study area would be covered by a parameter displaying a result derived from self-reinforcement.

In areas where there hasn't been an extensive amount of surveying, the implementation of this parameter would not be possible in the construction of a predictive model over said area, which makes it unusable.

Buffer zone distance intervals (metres)	Average Kvamme's gain	Average KS d-statistic
500	0,85	0,242
1000	0,837	0,72
1500	0,723	0,774
2000	0,595	0,782
2500	0,514	0,897

Figure 14: *The table displays the Kvamme's gain, and d-statistic values associated with the different levels of distance intervals. The green hue represents values which perform well either in predictive power or level of statistical significance, Dark hue represents distance intervals exceeding in either predictive power or statistical significance, while light hue represents lower, but still acceptable values in this regard.*

5.2.5. Slope and aspect:

The aspect variable was divided into two different categories: areas with south-, southwest- or southeast-facing slopes in one category and slopes facing the other directions in another. The category representing southward facing slopes scored a d-value of 0,118, which is slightly below the threshold and is therefore excluded from the analysis.

The reason for dividing the aspect-zones in broad and general categories is the necessity of generating analysis surfaces that can contain enough settlements and area cover. This means that in a study which covers the entirety of Scania, it may be possible to divide the entire surface in respect to each aspect category, while still maintaining the statistical significance of the result. This topic will be further examined later in this chapter, where the performance of the overlay is presented.

Neither slope nor aspect displayed any statistically significant relationship with settlement presence. The highest performance regarding the slope-variable is areas with slope gradients between 2 and 5

degrees, scoring a d-value of 0,040, which is below the threshold of 0,146 and is therefore excluded from the analysis.

5.3. Result

In this chapter I will go over the results from the analysis and answer the research questions based on the findings discovered there.

> Research question 1: "*Can a clustered pattern be observed, which would indicate the existence of influence from one or more environmental parameters?"*

Yes, the settlement distribution displays a clear pattern of clustering, which according to the Global Moran's I test of spatial autocorrelation is, with extremely high likelihood, not the result of random chance (figure 5). While the high degree of spatial autocorrelation doesn't explain the reasons behind the congregation of the features, it tells us one of two things:

- Either the surveying done in the area has been done almost exclusively around already known settlement sites, which results in a potentially misguided overview of settlement distribution
- Or there are indeed one or many environmental variables which are behind the clustered settlement pattern, resulting in the settlements grouping together.

It is much more likely that the second option is correct in this case. If the first option was correct, then the proximity to known settlements parameter would have outperformed any other parameter by a rather wide margin, including their overlay. As we will explore further down this chapter, this was not the case.

Research question 2: "*Based on the accessible data, Which variables are the most important for predicting the presence of prehistoric settlements in Scania? What are their relative levels of importance according to observed results and their performance together with each other?"*

The variables, which performs the best in a Scanian environment for archaeological predictive modelling purposes, with previous research in similar study areas in mind and through empirical testing are distance to modern coastline, distance to major lakes and rivers and on postglacial sand soil.

Their relative level of importance is (taking all buffer distances into account):

- 1. Distance to coast. This variable yielded an extremely high Kvamme's Gain value of 0,78 at a distance of 500 metres, with settlements being located very close to the modern shoreline.
- 2. Distance to major lakes and rivers. This variable yielded an extraordinarily high Gain value of 0,77 at the set distance of 100 metres.
- 3. Postglacial sand. Among the variables, which were able to display a statistically significant relationship with settlement presence, yielding a respectable Kvamme's Gain of 0,57.

Although both the distance to coast and lakes/rivers performed the best at the lowest set distances, the area covered by these are too small when overlaying them to judge the performance of the intersections between them, which is why other distances are used in this study. If a higher settlement distribution density was observed or the spatial scale was increased, then it would be possible to narrow down the size of the parameter cover areas, while still maintaining an acceptable level of statistical significance. At the chosen distances of 500 metres for major lakes and water bodies and 1500 metres for distance from coastline, the variables yielded the Kvamme's gain values of 0,54 and 0,66 respectively.

When spatially overlaid, the parameters perform extraordinarily well, yielding a Kvamme's Gain of 0,89. This further supports the validity of the suggested parameters to predict prehistoric settlement locations in the area. However, as I mentioned previously in this chapter, the smaller area a parameter covers and the smaller the sample, the more convincing proof we need to reject the null hypothesis of no correlation between the parameter and settlement presence. While the overlay of the parameters performs well in predicting settlement location with the area covering 1,45 % and containing 12,60% of all settlements within the study area, for this result to be statistically significant, a sample size of 87 is too small.

Following the formula of calculating the sample D-statistic for a sample size of 87 data entries ($0{,}146 = 1.358 / \sqrt{87}$), the difference between the percentage of contained settlements (12,60%) and the expected percentage (1,45%) does not exceed the calculated D-value, in which case we cannot assert statistical significance. For a parameter covering this percentage of area and containing this number of settlements, a sample size of at least 149 would be needed:

- Parameter d-value = 0,126 - 0,0045 (12,60% - 0,45%) = 0,1115
- We assign the result as the critical D, to get X, which is the lowest sample size required to exceed this value:
- $0{,}1115 = 1.358 / \sqrt{X}$
- $\sqrt{X} * 0{,}1115 = 1.358$ | *Multiplying both sides of the equation with* $\sqrt{X}$
- $\sqrt{X} = 1{,}358/0{,}1115$ | *Dividing both sides with 0,1115*
- $X = (1{,}358/0{,}1115)^2$ | *Exponentiating both sides to the power of 2*
- $X = 148{,}6$ | *Presenting the result, which is the lowest required sample size*

Based on this result, the conclusion can be made that if we want to narrow down areas of interest through the overlay-method when constructing predictive models, it would be desirable to have either a high settlement distribution density or a large sample, which could also mean larger study areas than the one presented here.

Although the correlation between the area overlaid by all parameters and settlement presence could not be claimed as statistically significant in this study, this should not be taken as a dismissal of the predictive capabilities of the parameters and their overlapping areas.

This only serves as a reminder that it would be preferable to have a larger sample to be certain that there is indeed a below 5% probability of the presented correlation being that of random chance, if we choose the standard threshold of $p = 0{,}05$ ($D = 0{,}1358$) as the uppermost value of the null hypothesis.

In summary, the parameters performed better in pairs than they did individually, which confirms the hypothesis that the overlay method yields greater results than only analysing the correlation between a single parameter and settlement presence. Interestingly, the intersections between two different parameters always performed significantly better than the combined surface of the two, with the intersection of all three performing the best (see figure 15).

	Percent of study area covered	Percent of all settlements in the study area	Kvamme's gain
All parameters collectively	43,20%	82,60%	0,48
Postglacial sand	18,10%	37,90%	0,52
Major lakes/rivers	23,60%	48,30%	0,51
Coastline	15,00%	43,70%	0,66
Either postglacial sand or close to major lakes/rivers	28,50%	47,10%	0,395
Overlay	5,00%	12,64%	0,6
Either postglacial sand or close to coastline	26,10%	60,10%	0,57
Overlay	5,50%	20,70%	0,73
Either close to coastline or major lakes/rivers	35,70%	78,20%	0,54
Overlay	3,00%	13,80%	0,78
Overlay of all parameters	1,45%	12,60%	0,89

Figure 15*: The table displays the performance of all parameters in pairs and all combined. "Overlay" refers to the areas where the paired parameters overlap. The performance is measured in Kvamme's gain, which tells us the degree of over- or underrepresentation of settlements that there is, given the extent of the area that is covered,*

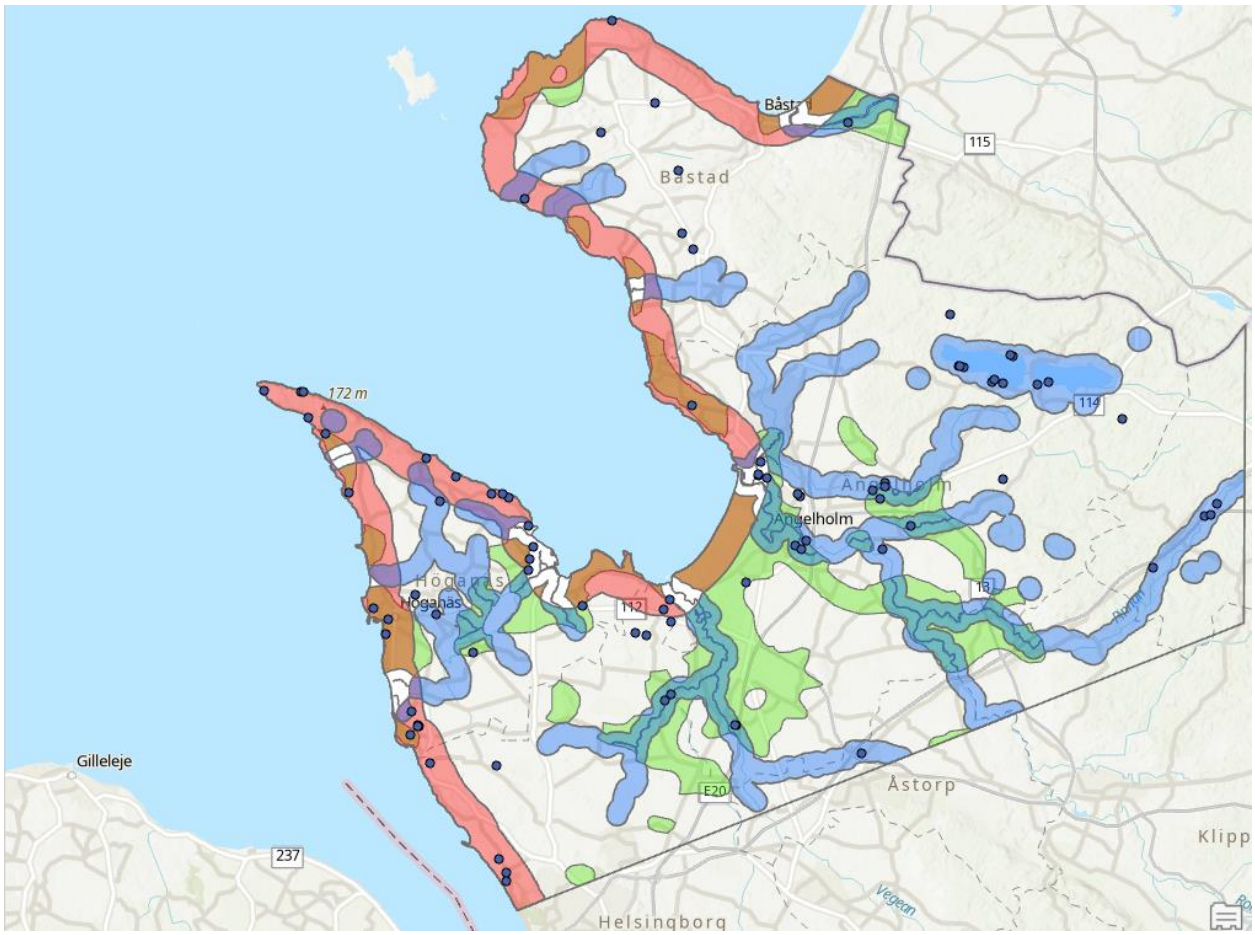


Figure 16: *The map is displaying the extent of all individual parameters along with the areas where they intersect. The areas that are 1500 metres from the coastline are displayed in red, postglacial sand is displayed in green and the 500-metre buffer zone around major lakes and rivers is blue. The intersecting areas are displayed by a combination of the colours of the overlapping parameters. The areas where all parameters are intersecting is displayed in white.*

> Research question 3: *Is it possible to extract fundamental principles and notions from this conclusion, which might aid the development of predictive models in other regions?*

Except for the affirmation of the expected performance of the tested variables for predicting settlement presence, the most interesting result from this study regarding archaeological predictive modelling is the relationship between parameter zone area, the percentage of settlements covered and the sample size. The major flaw in exclusively using the Kvamme's Gain value as the sole performance measurement is that it doesn't consider the size of the sample, which the Kolmogorov-Smirnov d-statistic covers up.
The relationship between the Gain and d-statistic values will be further explored in the discussion section below.

6. Discussion

In this chapter, I will reconnect with the different chapters of the paper to discuss the potential inferences and interpretations that can be made, as well as highlighting the shortcomings and limitations of the study and what role this paper has in the development of archaeological predictive modelling. This is proceeded by addressing the dynamic relationship between statistical significance and predictive power in the parameter performance evaluation. The chapter is concluded by suggesting improvements for future research and production of predictive models for archaeological use**.**

6.1. Reconnecting to the chapters

To summarise: In this study, an attempt has been made to evaluate the ability of a few chosen variables to predict prehistoric settlement locations in a Scanian environment, by examining a subsection of this region and testing the correlation between these variables and known settlement presence. To explain and warrant the methods used, a background of spatial analysis, the software environment and statistical techniques are given.

To get a starting point, which is grounded in previous research, variables that have been used successfully in similar areas are tested. This has the secondary effect of further increasing the confidence of using these variables in this or in similar environments, which enhances the suitability of using these variables for a predictive model covering other parts of, or the entirety of Scania.

The theory chapter covers multiple different necessary contexts. First, the place that predictive modelling has in modern archaeological theory is proposed through Kristiansen's "New paradigm". The second part covering the quantification of human spatial experience and resulting behaviour. The third part explains the theoretical framework behind spatial autocorrelation with the final part discussing the two different approaches to predictive modelling based on either data-driven induction or theory-driven deduction. This functions as a backdrop of what is presented in the methodology chapter, as well as illuminating the problems and potential in creating measurable parameters from the abilities of human senses.

To increase the applicability of these methods for other environments or future studies in the same region, the methodology chapter goes through the steps of evaluating the study area by first assessing the settlement distribution pattern and then the creation and testing of the parameters based on the variables. This is necessary to replicate and further test the methods for testing their scientific reliability and reproducibility.

In the materials and methods chapter, the features and data which will be subjected to the analysis is presented, with the parameter creation processing methods included. The parameters are created on a subjective basis and not observed features. For this reason, the process which has created them and the resulting objects are presented in the same section.

The analysis chapter goes through the implementation of the methods applied to the presented material, by analysing the correlation between settlement presence, the different parameters and the

intersections between them. After the performance of the parameters have been tested, the result is presented, which concludes the study.

6.2. Variable disclaimers

There are a few caveats regarding the result of this study, which will be discussed below.
While the result presented in this paper shows a very high degree of settlement overrepresentation within the parameters and their overlapping areas, it also exposes some limitations.

The variable “Distance to major lakes and rivers” is highly dependent on landscape reconstruction. As mentioned previously, the Scanian agricultural landscape has gone through significant changes throughout the 19th and 20th centuries, which has drained the rivers and marshlands. Thus, it is in many cases necessary to rely on historical maps or geographical features to recreate the prehistoric environment as closely as possible. The map used in this study from 1911, is certainly a better representation than what is visible on modern maps, but the settlements are in many cases from a period over 1000 years prior to the making of the map in question, during which period we can assume that the landscape has changed significantly.

The variable “Postglacial sand” is a rather specific kind of geographical feature, which is not common in all of Scania (see Helgesson 2002 p.7). In this particular case, this type performed the best among the soil types, but this may not be the case in areas where it is rarer. Due to the possibility of this soil type being evidence of past shorelines, it would make this variable the same as distance to coast at certain periods of time, which might explain its predictive power.

Which kinds of soil types are the most important for predicting settlement locations seems to depend on where the study area is located and whether the settlements are remnants of an agrarian society. Knowing these two factors, it is possible to form a reliable hypothesis of which kinds of soil types are potentially the best performing for predicting the location of ancient settlements. The variable of soil type seems to be reliable but choosing which kind of soil type to study seems to be highly varying from case to case.

The distance to coast parameter assumes that the coastline has been the same throughout the continuity of human habitation of the region. While this is not true, it is not possible to recreate a common shoreline for materials that are from entirely different time periods. Thus, this variable is highly dependent on high temporal resolution for recreating the conditions which were present at that moment. In areas where the shore is often several metres above sea level, generalising in this way is more acceptable. In the Scanian case, this may be an issue depending on how much the landscape has changed under the duration of the period studied. It is important to note that other regions are very reliant on high temporal resolution, such as the Netherlands, here it is much more necessary to be specific which time period is studied, because of the ever-changing shoreline. The issue of temporal resolution will be explored further down this chapter.

Another important aspect to consider is which kinds of variables are used in the construction of the predictive model. This varies to a high degree and could be described as a function of the region, what geographical features characterise the region and what culture that the material is remnants from.

To demonstrate this, I will give two examples of predictive models that were constructed at roughly the same time in widely different environments:

In cases where the topographical variance is high and the landscape is more arid, the relevance of variables such as soil type diminishes and slope gradient and aspect increases. One such example is the Piñon archaeological project in Colorado, USA. These variables are described as the most important environmental factors for predictability in the project (Kvamme 1992. pp. 25-28):

- Slope: too steep slopes interfere with human activity (negative correlation)
- Aspect: south-facing aspect yields greater warmth (positive correlation)
- Local Relief: high discrepancy in elevation between features close in proximity suggest rugged terrain (negative correlation)
- View Data: Locations that enables good visibility of the surroundings improves surveillance for the hunter gatherers (positive correlation)
- Shelter Index: a geometrical evaluation of how exposed the given location is to its surroundings (negative correlation)
- Distances to water sources: The constant need of water makes proximity to water sources a reliable predictor of human activity. In this example, horizontal as well as vertical proximity was taken into consideration due to the extreme terrain (negative correlation)

On the other hand, the researchers Roel Brandt, Kenneth Kvamme and Bert Groenewoudt conducted a research project in 1992, where the goal was to create a predictive model in the Regge valley of the eastern Netherlands akin to the models that were created in the United States at that point in time. The region where the study area is located, the topographical elevation variance is very low. This makes the importance of factors such as visibility from high altitudes and slope gradients deemed unusable by the group of researchers (Brandt et al 1992. p. 3). It is worth noting that this study was carried out without the access to LIDAR data, which could provide a more detailed DEM for basing the slope variable on.

The variables studied in the Regge valley project are proximity to water sources, geomorphological type, soil type/texture. The researchers mean that it would have been appropriate to include socio-cultural factors in this case, like proximity to roads and other settlements, although these were not included due to the scarce amount of available data (Brandt et al 1992. pp. 5-6).

The Regge-valley project is a good example of a predictive model constructed for a flat landscape, where topographical variance is low and a long history of continuous settlement of agricultural societies exist. The researchers acknowledged the fact that their model would have to be based on very different environmental variables than what has been chosen in the models created in the western parts of the United States previously.

This illustrates that a single model consisting of a certain set of parameters cannot be assumed to be applicable in all environments, despite how well it performs in one or more particular environments.

For this reason, I do not suggest that using the result from this study as the sole basis for the construction of a predictive model over the region of Scania is to be done. Instead, it is an attempt to test the variables, which could potentially be included in the model. To be able to validate this result,

a more consistent framework where the parameters are created would be needed. This applies mainly to the variable of "major lakes and rivers", which are drawn in the ArcGIS Pro software based on a map from 1911 and therefore is very specific for this particular study area.

6.3. Temporal resolution

Temporal resolution is an issue which affects both scientific validity and predictive power of a model. The range of potentially different time periods that the studied settlement material might be from would ideally be narrow, although this comes with a few problems:

- It may limit the knowledge we can gain about the ancient world:
  If the resolution is too low (no separation between time periods or few defined chronological categories), then we cannot be sure how settlement distribution varies depending on the time-period.
- It may limit the possibility of conducting a statistical analysis:
  If the resolution is too high (many chronologically separated categories), the sample might contain too few data points per category, and we are as a result unable to claim statistical significance to the results we acquire.

As the study done by Bo Ejstrud shows (Ejstrud 2003), a predictive model created by analysing material with clearly defined time periods associated with them can enlighten us about the settlement distribution specifically for these time periods, which makes the model a great tool for understanding the social landscape of the ancient world and its evolution, not only for assessing the probability of encountering archaeological material in an area.

An important factor to keep in mind is the changing landscape over the millennia. If we have a large sample of archaeological material spanning a few thousand years, then we can expect to see a great difference in natural features such as shoreline and extent/shape of rivers and lakes depending on the time-period, as well as the expected importance of soil types, which is dependent on the prevalence of agricultural activity.

It's therefore important to at least consider the possibilities of distinguishing between material from different time periods if the sample size is adequate for creating a model with the material divided in different chronological categories. If this distinction is made, then it might be appropriate to also consider the changing landscape and the human interaction with it over the span of the collected material's lifetime.

6.4. The issue of data quality

While the quality of the data presented as tested in this study has been high in general, there are some issues that needs to be addressed.
The soil type layer used in this study, provided by the Geological Survey of Sweden (SGU), has a resolution of 1:000 000, which is appropriate for visualising the distribution of different soil types on national level, but it might not be appropriate for analyses on regional level or lower. This is explicitly stated in the shapefile documentation by SGU**.** Although using this variable showed promising results this time, as it has shown in previous studies, using a dataset with resolution this

low might diminish the predictive power of soil types covering small areas to the advantage of soil types covering large areas, due to excessive generalisation.

Although very useful for these kinds of studies, the data containing the information of settlement locations provided by Riksantikvarieämbetet do not include distinction between time periods, nor a clear account of whether any two settlement location points that are close to each other, are separate entities or in fact the same settlement but at two different locations. This can be a major issue in archaeological predictive modelling, due to the importance of settlement density when assessing the performance of the parameters. The clearer temporal and spatial distinction made within the site-dataset, the more accurate parameter performance assessment we can expect to make.

One of the major issues encountered while conducting this study is the slow processing speed when attempting to create a digital elevation model based on LIDAR data points (1x1m). For large study areas at a Swedish sub-state/multiple county level, it might be advantageous to use either a lower spatial resolution on the LIDAR dataset or using hardware with great processing capabilities to create a detailed triangulated elevation surface. A surface displaying slope degrees in high detail may be critical to determine the importance of this variable in areas with low slope degree variance such as Scania.

6.5. Predictive power and statistical significance

What the result has shown is that including the Kolmogorov-Smirnov test could further enhance the evaluation of parameter performance. For the result of predictive models to be reliable, the variables of which it is based on, needs to have been tested with a sufficient sample size. This also means that the area, which a parameter covers, either should contain a high density of sites or cover a sufficient extent to contain a certain sample size of sites. Predictive power as a measurement derived from the gain value, does not discriminate between results from a small or large sample, nor does it if the model area cover is small or great.

The dependency on large samples to yield statistically significant results, is a solid argument for why both parameter testing and predictive modelling may be preferrable to do on a large scale, alternatively on a small scale with high settlement density.

In summary, the relationship between predictive power and significance is highly dependent on the sample size, with larger samples representing a more accurate picture of predictive power due to the lesser probability of the result being that of random chance.

6.6. Suggestions on future research

The result presented in this paper and what contribution it provides, is a small part of a long series of methodological validation regarding prehistoric settlement distribution with respect to geographical properties. Its role in archaeological predictive modelling is to further test the established methods and provide suggestions on further improvement. Continuously testing the methods applied to test the correlation between human settlement presence and environmental variables are essential for both developing these methods and enhancing our understanding of the past, by making more

accurate predictions than were previously possible. In this subchapter, several suggestions on future research will be presented to improve this process.

- By utilising the possibilities of prehistoric landscape and ecology reconstruction, we could potentially recreate the settings which influenced the human spatial behaviour we are observing. If this process was to be carried out for an entire study area, it could greatly improve the ability to predict settlement presence.

- Unbiased surveying, completely ignoring any environmental variables while identifying settlement locations in an area that has not yet been surveyed, could be very useful for model validation. This would preferably be done by dividing the area in regularly sized sections and randomly selecting a number of these for surveying. In this way, the results from the correlations tests would not be skewed by already taking the variables into account during the survey.

- Compare the performance between different statistical techniques and observe the results. The technique which proves to be performing better than the others on a consistent basis could be the desirable option to use when a predictive model is created.

- It could be useful to compare the spatial distribution of settlements and the distribution of the variables. The law of autocorrelation suggests that a naturally occurring observable feature should not be randomly distributed, but if exceptions to this rule were found, it could invalidate the testing of the degree of settlement clustering as a method for suggesting the presence of inferring external variables.

- Explore the potential of AI and machine learning in identifying suitable conditions for settlement presence. By developing the methods and techniques which are used when constructing predictive models to a point where they are consistently performing well on a regular basis, they could be integrated into computer algorithms and the computer could do the analysis for us if it has access to the necessary material. This could have great potential for the creation and testing of models on many study areas, much more efficiently than what could be done by a human agent. This could for example be done in the programming language Python, which is integrated into the commonly used ESRI ArcGIS software.

7. Conclusion

Archaeological predictive models have many different facets, and therefore a great number of areas which can be developed and further improved upon. The progress which has been seen in prehistoric human spatial behaviour theory, as well as GIS technology, statistical techniques and even the potential of machine learning implementations in the last decades, are promising signs for this area within archaeology.

The Scanian environment, with its relatively large quantity of registered settlement sites, is a great testing ground for the development of predictive models in the future, which this study is a steppingstone for.

8. References

Asserstam, Marcus. "*Predicting mesolithic pioneer settlements in eastern middle Sweden*" 2010.

Berglund, Björn E. & Rapp, Anders ``*Geomorphology, climate and vegetation in north-west Scania, Sweden, during the late Weichselian"* in Geographica Polonica. 1988

Brandt, R., Groenewoudt, B.J. and Kvamme, K.L. *An experiment in archaeological site location: modelling in the Netherlands using GIS techniques*. 1992

Chapman, Henry "*Landscape Archaeology and GIS"* 2006

Ducke, Benjamin "*Regional Scale Predictive Modelling in North-Eastern Germany*" 2010

Ejstrud, Bo "*Indicative models in landscape management: Testing the methods"* 2003

Gillings, Mark "*Landscape Phenomenology, GIS and the Role of Affordance*" 2012

Helgesson, Bertil "*Järnålderns Skåne: samhälle, centra och regioner*". 2002

Jones, Eric E "*Using Viewshed Analysis to Explore Settlement Choice: A Case Study of the Onondaga Iroquois*" 2006

Judge, William J. and Lynne Sebastian. "*Quantifying the Present and Predicting the Past: Theory, Method, and Application of Archaeological Predictive Modelling"*. 1988

Kristiansen, Kristian "*TOWARDS A NEW PARADIGM? The Third Science Revolution and its Possible Consequences in Archaeology"* 2014

Kvamme, Kenneth. L. *A predictive site location model on the High Plains: An example with an independent test*. 1992

Kvamme, Kenneth. L. "*One-Sample Tests in Regional Archaeological Analysis: New Possibilities through Computer Technology*". 1990

Kvamme, Kenneth. L. "*Spatial Autocorrelation and the Classic Maya Collapse revisited: Refined techniques and new conclusions"* 1990

Lee, Peter M. "*Statistical tables*" 2005.

Lock, Gary., Kormann, Mariza., and Pouncett, John. "*Visibility and movement: towards a GIS-based integrated approach"* 2014

Lockhart, Richard "*Unbiased estimation*" 2020

Löwenborg, Daniel "*Excavating the Digital Landscape: GIS analyses of social relations in central Sweden in the 1st millennium AD".* 2010

Massey, Frank J. "*The Kolmogorov-Smirnov Test for Goodness of Fit*". 1951

Mihu-Pintilie, Alin & Nicu, Ionut Cristi. "*GIS-based Landform Classification of Eneolithic Archaeological Sites in the Plateau-plain Transition Zone (NE Romania): Habitation Practices vs. Flood Hazard Perception*". 2019

Nicu, Ionut Cristi , Mihu-Pintilie, Alin & Williamson, James. "*GIS-Based and Statistical Approaches in Archaeological Predictive Modelling (NE Romania)*". 2019

Nsanziyera, Ange Felix. Rhinane, Hassan. Ouja Aicha and Mubea Kenneth. "*GIS and Remote-Sensing Application in Archaeological Site Mapping in the Awsard Area (Morocco)*" 2018

Ortman. Scott G, Varien. Mark D & Gripp. T. Lee. *"Empirical Bayesian Methods for Archaeological Survey Data: An Application from the Mesa Verde Region"* 2007

Otarola-Castillo, Erik & Torquato, Melissa G. *Bayesian statistics in archaeology.* 2018

Påsse, Tore & Daniels, Johan. "*Past shore-level and sea-level displacements"* 2015
Schrader, Lucian Norman "*Demonstrating GIS spatial analysis techniques in a prehistoric mortuary analysis: A case study in the Napa Valley, California*" 2013

Svedjemo Gustaf, *Predictive model for iron age settlements on Gotland, 200-600 AD*. 2003

Van Leusen, Martijn. *Predictive modelling for archaeological heritage management: a research agenda*. 2005

Van pool, Todd L. & Leonard, Robert D. "Quantitative analysis in archaeology". 2011

Vaz, Eric "*Archaeological Sites in Small Towns—A Sustainability Assessment of Northumberland County*" 2020

Verhagen, P. *Case studies in archaeological predictive modelling* (Vol. 14). 2007

P. Verhagen, H. Kamermans, M. van Leusen and B. Ducke. *New developments in archaeological predictive modelling*. 2010

Verhagen, P. Predictive modelling. *The Encyclopaedia of Archaeological Sciences.* 2018

Waters, Nigel "*Tobler's First Law of Geography"* 2017

Wheatley, David, and Mark Gillings. *Spatial Technology and Archaeology: The Archaeological Applications of GIS*. 2002

Yaworsky. Peter M, Vernon. Kenneth B, Spangler. Jerry D, Brewer. Simon C & Codding. Brian F. "*Advancing predictive modelling in archaeology: An evaluation of regression and machine learning methods on the Grand Staircase-Escalante National Monument".* 2020

8.1. Digital sources:

URL: https://www.sgu.se/om-geologi/jord/fran-istid-till-nutid/landhojning-fran-havsbotten-till-lerslatt/postglacial-sand-och-grus/’

Stephanie Glen. "Moran’s I: Definition, Examples" From StatisticsHowTo.com**:** Elementary Statistics for the rest of us! viewed 17.3.2022. URL: https://www.statisticshowto.com/morans-i/

Spatial Autocorrelation (Global Moran’s I) (Spatial Statistics) viewed 18.3.2022. URL: https://pro.arcgis.com/en/pro-app/2.8/tool-reference/spatial-statistics/spatial-autocorrelation.htm