Predictive Model for Iron Age Settlements on Gotland, 200 – 600 AD.
- An overview of Predictive Modelling in Archaeology with an implementation of Logistic Regression Analysis

By Gustaf Svedjemo
Högskolan på Gotland (Gotland University), SE-621 67, Visby, Sweden
Department of Computer and Information Science (IDA)
Linköping University, SE-581 83 Linköping, Sweden
2003

Abstract
Predictive modelling in GIS has been in practise in the USA for several decades. Lagging some years behind it has been tried in some European countries. This paper briefly outlines the concept of predicting modelling and describes an implementation of it made on the island of Gotland, Sweden. The method used is Logistic Regression Analysis and the object of the study is the Iron Age settlements, Kämpegravar. The independent variables used to predict the areas are soil types and land use and settlement patterns in the 18th century, derived from old cadastral maps.

Table of Contents

Table of Contents .................................................................................................. 1
Introduction ........................................................................................................... 2
1 Predictive Modelling in Archaeology .............................................................. 2
  1.1 Information, methods and outcome .............................................................. 5
    1.1.1 Information ............................................................................................. 5
    1.1.2 Methods ................................................................................................. 6
    1.1.3 Outcome ................................................................................................ 8
2 Case study ......................................................................................................... 9
  2.1 Objectives .................................................................................................... 9
  2.2 Method ......................................................................................................... 9
  2.3 Dependent Variable and Unit of Study ....................................................... 10
  2.4 Model Areas ............................................................................................... 12
    2.4.1 Training Area ....................................................................................... 12
    2.4.2 Verification Area ................................................................................... 13
    2.4.3 Prediction Area .................................................................................... 13
  2.5 Independent Variables ............................................................................... 13
  2.6 Model Development ................................................................................... 18
  2.7 Results ....................................................................................................... 20
  2.8 Conclusions and Discussion ...................................................................... 22
References ......................................................................................................... 24
Introduction

The need and demand for ways to predict and understand the reasons for the location of various archaeological sites is of great importance, both for Cultural Resource Management (CRM) and for Academic research in Archaeology. Predictive modelling is one of the first more complex operations that were done in GIS by archaeologists. It was the natural continuation of other, non-computer based, techniques developed to be able to predict sites in North America. Very few inventories for archaeological sites had been conducted and there were a need to develop methods to predict the areas of archaeological sensitivity were, when new land uses were planned (Altschul 1990, Della Bona 1994). Lagging some years behind, European CRM-agencies and scholars has begun with Predictive modelling. Since the early 90’s a nationwide project is set up in the Netherlands to find models for predicting Archaeological sites (Kamermans & Wansleeben1999, Deeben et al 1997). The European attempts have not, in most cases, reached the maturity of the American efforts. The Europeans are still finding their way into this technique (Stančič & Kvamme 1999). One reason for the early start in America is the greater need of predictive models. Another reason is the difference in legislation. In most European countries, archaeologists can survey and excavate on privately owned land. This is not the case in America, were archaeologists most often are confined to publicly owned land (Stančič et al 2000).

To the authors knowledge no efforts have been made in Sweden, at least not published, to use GIS in predictive modelling. This paper will make a first effort to test predictive modelling on the island of Gotland in Sweden.

1 Predictive Modelling in Archaeology

A predictive model attempts to predict were archaeological sites or features are located, by looking for tendencies and patterns observed in a region or by theory and notions of the distribution of sites or features (Kvamme 1990). CRM and Academic use normally need two different approaches. In CRM the main focus is to be able to predict where it is likely to encounter Archaeological sites in an area. The need to explain why it is there is of minor interest. In Academic research the need for understanding the reasons and mechanisms behind the choosing of a site location is often the central issue. The former is predictive in its aims and the latter is interpretive (Warren 1990 a, Leusen, van 1996). The Predictive approach can be expressed as looking for locational factors and the Interpretive as understanding locational choice factors. The difference can also be expressed as that the predictive case one predicts the present distribution of sites and in the interpretive case one tries to explain the past distribution of sites (Leusen, van 1996). It can also be viewed as the difference of using deduction from theory or induction from observations (Warren 1990 a, Della Bona 1994). This would be equivalent to what van Leusen calls the Academic (deductive) and CRM (inductive) approaches. Kohler and Parker call them empirical correlative and deductive Models. The former is based on
observations and the latter is based on deducted theory (Kvamme 1990). Basically it is the same as the CRM and Academic strategies. The different approaches are, in some sense, also the difference between how these techniques are implemented in North American and European Archaeology. The predicting approach, is prevailing in the North America and the more Interpretive, is more common in Europe. This is due to different traditions and the difference between the archaeological material and knowledge in the two areas (Harris & Lock 1995).

In practice the boundary between the two approaches is somewhat fuzzy. Most predictive modelling is a cyclic process, were both deductive and inductive input is used, to refine the model (Kamermans & Wansleeben1999). Warren means that the theories are needed in choosing the independent variables. If only observations are used in the variable selections, there is a risk for the creation of a weak model (Warren 1990 a).

In this paper the term Predictive Model will be used for both approaches, and only if there is a need to distinguish the two, will it be explained which approach is meant.

In the pure predictive case there is a danger, that when no attempt is made to understand the underlying reasons and structures for the location, a biased result will be obtained. This is due to the fact that the archaeological material itself is biased. In the interpretive approach, which theoretical base was formed by the scholars of New Archaeology in the 60’s and 70’s, challenges a huge effort in reconstructing all the relevant factors in past societies (Leusen, van 1996).

Most predictive models use environmental variables that are measurable in the present landscape. This implies that there are at least two general assumptions. The first is that environmental factors played a key role when the prehistoric people choose their place to settle and the second is that these variables are still measurable in the present landscape (Warren 1990 b).

The almost exclusive use of environmental variables has lead to criticism of predictive modelling as being to environmental deterministic (Leusen, van 1996; Harris & Lock 1995). This is in the majority of studies true. This can be due to the fact that cultural factors are harder to map and thus getting in to a GIS. Some analysis has been made that has tried to take human cultural behaviour in account, with the use of viewshed, friction surfaces etc. But as Kvamme and van Leusen points out, they are all derived from the environmental variables, like DEM’s, and thus suffer the same limitations. The problem is that in a GIS it is the patterns that can be analysed, not the random occurrences. The GIS can identify the non-pattern parts, but not analyse them (Gaffney & Leusen, van 1995).

Wise (2000) emphases the difference between ecological and phenomenological variables to be used in GIS. Most GIS applications in Archaeology today are solely using ecological variables, and do so in the wrong way. She means that GIS is a perfect tool also for phenomenological variables. In her paper she outlines some ways to incorporate human perception and consciousness of the landscape and environment into GIS analysis.
Another criticism of the use of predictive modelling is that the focus is often on the wrong things. For most model builders it’s often solely the predictive power of the model that is important and they do not use the full power of the models. Altschul (1990) points out that predictive power is very often poor, but still there are much information and knowledge to extract from the results, if you focus on the right things. In his study over Mount Trumbull he analyse the outliers, what he calls “Red flags”, which are the sites that was not predicted right. From these studies he can derive much knowledge about the studied ancient societies (Altschul 1990).

Below is the model process, as outlined by Della Bona (1994). Most predictive models just conduct the two first stages. But with Altschul’s ‘red flag’ model, the process continues, by analysing also the outliers (The sites that was not predicted by the initial model) and incorporate new information in the model and thus refining it. This is of course a time consuming and expensive process, which can explain why it is not often performed.

<table>
<thead>
<tr>
<th>Primary Stage</th>
<th>Hypothesis building - data collection strategies</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Initial data collection</td>
</tr>
<tr>
<td></td>
<td>Field reconnaissance, collection of baseline data</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Secondary Stage</th>
<th>Deductive phase of modelling</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Association between environmental variables and sites</td>
</tr>
<tr>
<td></td>
<td>Literature review and integration into model(s)</td>
</tr>
<tr>
<td></td>
<td>Development or application of initial model(s)</td>
</tr>
<tr>
<td></td>
<td>Testing of model(s) on previously surveyed areas</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Tertiary Stage</th>
<th>Continued application of model(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>New data continuously incorporated into process</td>
</tr>
<tr>
<td></td>
<td>New sites discovered are interpreted and incorporated into existing model(s)</td>
</tr>
</tbody>
</table>

Table 1. Summary of the three stage modelling process (from Della Bona 1994).

To produce a successful predictive map one needs to have several models for different site types. The preferences of site location have change over time and reasons for the site (Dalla Bona 1994, Stančič et al 2000). The different trades of the hunter-gatherers, Iron-age farmers, fishers etc lead to different strategies for choosing sites. At best one model can account for a couple of different kinds of sites, and related features. In American predictive modelling it seems not to be so common to take this fact in account. It’s seems more to be the case in the European examples that I have studied.
An American example of it is Hasenstab and Resnick (1990) when working with both historic sites (farmsteads, mills, trash dumps) and pre-historic sites. They failed to predict Maple sugar mills, because the location of these were governed by completely different variables than e.g. the farmsteads. Another example of a study that would have required several models is the study Stančič et al (2000) conducted on the Island of Brač. The regression analysis they performed gave a poor result and it was determined that it was the fact that the archaeological material consisted of several different types of settlements, that would require a separate model for each type.

The complex situation of Gotland’s ancient remains makes predictive modelling very hard. To get a full predictive model for all kinds of ancient remains on Gotland would take a great deal of effort and a vast number of different models.

Predictive modelling in GIS is not only a matter of predicting sites and locational analysis in raster or vector systems. Predictive modelling can also use other techniques like network theory and analysis to predict settlement spread along waterways, like Zubrow (1990) did in a study of migration and settlement spread in New York state. This paper however, will only deal with, what we can call, traditional site locational predictive modelling in GIS.

1.1 Information, methods and outcome

All predictive models consist of three basic elements: Information, method and outcome. A predictive model uses a method to transform information into a prediction, outcome (Warren 1990 a).

The unit of study is not the archaeological site, but rather the land parcels (unit) themselves (Kvamme 1988 b). The aim of the model is not to study the archaeological sites or features, but rather to predict if a site is present in a unit or not, or to interpret why it is there.

1.1.1 Information

The information used in the models is of two parts, the dependent and the independent variables. The dependent variables are the archaeological sites or features whose distribution is sought. There are several kinds of dependent variables to use (from Kvamme 1990):

1. *Site presence or site absence*. This is a binary technique that deals only with dichotomous dependent variables. It predicts the presence or absence of sites in a land parcel. This is the most frequently used technique.
2. *Multiple site types*. A polychotomous model, which also can deal with many different site types and can predict the likelihood of the land parcel containing a site of a specific class or no site at all.
3. **Counts or site density.** These are models normally used for studies where the study unit (land parcel) is large. The frequency or density of sites or features is calculated.

4. **Site significance.** By ranking different sites by significance these models attempt to rank locations after their archaeological importance. The initial ranking is done by theory.

The independent variable is the characteristics that are recorded at each land parcel. These characteristics can be divided in four major themes (from Kvamme 1990):

1. **Environmental variables** that can be measured, like elevation, precipitation, vegetation, soil type etc. at every unit. From some of these variables others can be derived, e.g. a DEM can be used to calculate the slope, aspect, relief and landforms, and from a stream the distance can be calculated.

2. **Cultural and social factors** can be independent variables. These variables can be roads, other settlements, central places etc. Often they are derived from environmental variables, like viewshed, friction surfaces etc.

3. **Positional characteristics.** These variables are based on autocorrelation. If clustering, dispersal etc are observed, it can be used in the model. Models are also developed that can predict the coordinates of site locations.

4. **Radiometric characteristics.** These variables use the measured reflectivity of radiation in the unit. The raw data is most often remotely sensed data like Satellite images and Aerial photos. Ground cover radiation is however, in most senses, just a proxy for environmental variables present in the unit.

**1.1.2 Methods**

The methods equal too which models and techniques are used in the prediction. The models/methods can also be described as a set of decision rules (Kvamme 1990). There are several methods to use when making a predictive model in a GIS environment. They normally implement some kind of statistical method. It is in the field of Predictive Modelling in the North America that spatial statistics in archaeology begun (Harris & Lock 1995). One of the leading names is Kenneth Kvamme, who have adopted and integrated several advanced statistical methods for use in GIS and Archaeology.

The methods can be grouped in two major categories. Methods based on trends in location only and methods based on trends in locational characteristics. The former one deals only with the positional (x,y) coordinates as independent variables. No other information is used. The latter, which is the most common one, uses different characteristics measured at the sites (compared to non-sites) to establish the trends in location.

There are a vast series of methods and model types. I will only mention some of the more common ones. For a full description of them all is beyond the scope of this paper. For a description of many of them, see Kvamme (1988 b, 1990), Della Bona (1994) or Judge & Sebastian (eds. 1988).
**Boolean models.** This a very straight forward approach, that is easy to implement without any deeper knowledge in statistics. It can also be used when very little sample data is available. For each variable a threshold value is calculated and then stored in separate Boolean layers, where 1:s meet the threshold and 0:s don’t. The layers are then summed and areas meeting all criteria’s, and thus are most likely to contain sites, are given the highest value (= number of variables) and the fewer criteria’s are meet, the lower the number. There are of course drawbacks attached to the simplicity. The main drawback is that you can’t estimate the relative influence of each variable, since the variables are treated separately. You can however assess the performance of each variable by calculate how large area it drops of the study area (Stančič & Kvamme 1999).

An example of work using this model is Stančič’s and Kvamme’s predictions of Bronze Age hill forts sites on the Island of Brač, Croatia, where mostly environmental variables were used (Stančič & Kvamme 1999).

In his study of Etruscans and Roman settlements in Albegna Valley, Italy Perkins (2000) uses this approach. Since the sites range from a long period in time and different cultures, several models were calculated. The variables used in the different models were, elevation, slope aspect and soil types.

In their survey over Fort Drum Hasenstab and Resnick (1990) worked with both pre-historic and historic sites. The historic sites were farmsteads and mills from the 19th century the mid 20th century. They worked with two different models, one for each period. In the historical model, historical maps from the late 19th century were an important source. From the maps some historical sites could be identified and the old road network, both were chosen as variables. Map layers of distances to water, roads, farmsteads and soil types built the model that was based on simple map overlays. The prediction maps produced, showed of low value in the field and the model was abandoned. And a new strategy was introduced. The whole project area was surveyed with a walkover. After the survey was conducted the different sites identified was analysed for non-random tendencies using Discriminant analysis (Hasenstab and Resnick 1990). This can be seem as an example of how low the prediction power can be for simpler models.

**Weighted map layer models.** This is basically the same approach as the Boolean model, but instead of just assigning 1:s and 0:s to variables, that either meet or do not meet your defined threshold, you have several thresholds and weight them. The main drawback with this method is that you get very different results, depending how you define your weights. There is no rule or standard procedure for this.

One of the first attempts in Dutch archaeology with predictive modelling used this approach. They solved the weighting problem, by using an empirical method and simple statistics to calculate the weights (Brandt et al 1992).

**Regression / correlation models.** There are a vast number of different methods based on correlation between the independent and the dependent variable. Most statistical methods
need a relatively large number of samples (> 25) and are parametric, which means they have some underlying assumptions of how the data is distributed. For example, in many cases, like Linear Regression and Discriminant Function Analysis, there is an assumption that the independent variables are normally distributed. This is often hard to meet in an Archaeological context, and leads to more work in transforming the data to meet these assumptions. In practice however, there are many examples of models performing well, without the underlying assumptions are met (Kvamme 1990).

An example of the use of linear regression analysis is Stančič and Veljanovski (2000). They worked in the same area as Stančič & Kvamme (1999), but this time with Roman settlements.

Discriminant Function Analysis is a widely used method, which for example was used in the Kromme Rijn Area in the Netherlands (Deeben et al 1997), was it showed a good predictive power, with the variables used. More then 75% of the settlements were predicted in the zones with the highest likelihood. These areas only covered 25% of the model area.

Logistic Regression Analysis is a regression technique, which is nonparametric, and thus don’t have any such assumptions. It has been widely used in many models. It will be described more thoroughly later on in the case study chapter, section 2.2.

1.1.3 Outcome
The outcome of the models can be at different scales of measurement: nominal, ordinal, interval or ratio. For Archaeological site location an output in a ratio scale, which means that a prediction of probability can be produced, is to be preferred (Warren 1990 a). Logistic Regression Analysis produces a probability surface. The output of various overlay methods produces an outcome on the interval or ordinal scale, the higher the value, the more of the used variables is present.
2 Case study

2.1 Objectives

As stated earlier the main objective of this study is to see if a predictive model for Iron Age settlements can be produced for Gotland. This would be very usable for the main agricultural districts of the island, were the farming processes have removed most such remains and the landscape is much altered from the one that existed during the Iron Age. All the large bogs and wetlands are drained and streams are altered. Nearly all land is cultivated and if you see the landscape today, it is very hard to picture the Iron Age environment.

A second objective is to see if the old cadastral maps can be used according to Dan Carlsson’s results, since the variables most commonly used in Predictive modelling are probably not usable on Gotland. This will be further discussed below.

2.2 Method

The method chosen was Logistic Regression Analysis (LRA). This is a statistical method that is quite frequently used and a standard method in many fields of analysis. It is considered one of the best for Archaeological predictive modelling. It produces a probability map in the ratio scale. The biggest drawback for most archaeologists is that it is quite complex. However, nowadays it’s built in as standard module in many GIS-applications like Idrisi32 and ArchInfo, and is also available in many Statistical packages for standard PC’s. Kvamme introduced the method into Predictive modelling in Archaeology in 1983. The main advantages with LRA are that it can deal with all scales, nominal, ordinal, interval and ratio and don’t assume the data to be normally distributed. This means that you can use dichotomous dependent variables, like presence/absence of sites in a unit and use soil classes’ (nominal) together with distances (ratio) in the same model.

The general formula for logistic regression is:

$$\text{Probability (event)} = \frac{1}{1 + e^{-z}}$$

Where

- $z$ is $b_0 + b_1x_1 + b_2x_2 + \ldots + b_px_p$ (The logistically derived discriminant function)
- $e$ is Natural logarithm
- $b_0$ is The intercept
- $b_1, b_2, \ldots, b_p$ are The coefficient/estimates
- $x_1, x_2, \ldots, x_p$ are The variables/parameters
In his model for the Shawnee area in southern Illinois Warren (1990 b) uses logistic regression analysis (LRA) in his model. Twenty-six environmental variables were tested in the model. Eight of these were selected as strong indicators by the LRA and included in the final model. The results from the model showed that it was not as good as expected. At best the model could predict 67% of the site locations, but only 39% of the non-site locations. Warren means that the models general tendencies are valid. The main problem was the low quality of the DEM.

Another problem Warren (1990 b) identified is of patterned residuals. Some variables showed a bimodal density and the LRA he used was a linear one, which led to a systematic error. Warren means that this easily could be overcome by defining the variables as non-linear instead.

Thirdly Warren (1990 b) identifies a problem that is common in many statistical methods. That is the problem of rare or poorly represented nominal-scale variables. They have a tendency of behaving unstable. One solution can be to identify the weak variables and combine them to create stronger ones.

His conclusions are that the quality of the data must be carefully evaluated. Furthermore he emphasises the need to analyse the variables before they are entered into the model, so that weak variables and other properties, like bimodality, can be identified (Warren 1990 b).

Stančič et al (2000), as mentioned above, used a regression model in their study of Roman settlements on the island of Brač.

In a similar manner as Warren (1990 b), Carmichael (1990) used Logistic Regression Analysis and univariate analysis of the different variables in his study of site locations in north-central Montana. He uses fewer independent variables, but the variables are shown in previous studies by Kvamme to be relevant for site location analysis of Hunter-Gatherer sites. He reaches a higher number of correct classifications; 72% of the sites and 55% of non-sites.

2.3 Dependent Variable and Unit of Study

The sites that are in focus in this study are dating from the period 200 AD – 600 AD. They consist of the remains of the farmhouses. In folklore they are called “Kämpegrav”, which means giants grave. It was not until late 19th century that it was established that they were the remains of houses. They will henceforth be referred to as Iron Age houses. This is the only prehistoric period in Gotland that stone houses were erected. They are of a simple kind with quite low dry masonry walls. They can also be found on the neighbouring island of Öland and in parts of Östergötland, on the Swedish mainland. They are believed to be the absolutely predominant house type for farmsteads during the actual period (Carlsson 1979). In the main agricultural districts of the island were the
modern agricultural activity is very heavy, there are very few. It is believed that there have been many farmsteads there during the Iron Age, but the evidences are no longer visible in the ground surface. There are about 1,800 remaining visible in the landscape on Gotland. Dan Carlsson (1979) estimates it to be around 40% missing, and that total number was around 800 farmsteads, if an average of 2.9 houses per farmstead is used. Based on other data, Östergren (1989) comes to a different result of around 2,253 to be the total number of stone house foundations and that 60% are missing today. There have been several estimations done over how many stone houses there were, at the time of their use, for an overview of the discussion see, Österberg 1989 and Carlsson 1979.

Linked to the Iron Age houses is also another type of ancient remain, called ‘stensträngar’. These stonewalls were used to fence the infields, thus preventing the livestock and wild animals from grazing in the fields. Many of these still remain in the landscape. They consist of low stonewalls, made of relatively big stones (often around 0.5 meter in diameter) and can be several hundred meters in length. It would seem natural to include them in a prediction model, but they were decided to be left out. The reasons for this are several. Firstly they are absent in the same area as the houses, thus making them poor predictors. Secondly, the fencing systems can be very extensive. In the Ancient remains database (fornminnesinventeringen) they are recorded with just a single coordinate point, often together with the house remains. Thirdly, the database contains no dating and professional judgements of the remains, only an ‘objective’ description of the remain, thus making it hard to judge whether they are of Iron Age dating or not. This is a common problem with many archaeological databases (Kvamme 1988). The vast majority are Iron Age, but some are located in places that are very doubtful (see fig 1). Within the scope of this paper there are no means to adjust these things, so they are left out of the model.

Since the object under study is a farmstead, the unit of study can be many. There are several ways to attack this problem. One way would be to look at the entire farmstead, including the infields. This is the approach chosen by Deeben et al (1997) when facing a similar problem. In their model for Roman period settlements in the Kromme Rijn Area, the Netherlands, they decided on a simplification. A circular area of 400 meters around the central point (Hamlet), which contained the infields, was considered an ‘Archaeological relict area’. This was their unit of study. The same approach could have been chosen here, but since the farmsteads vary much in size and are hard to estimate and the fact that the houses were situated on the rim of the infields, it was decided just to use the location of the houses.

The locations of the Iron Age settlements are found in the Ancient remains database kept by the County museum, Gotlands Fornsal, which is a locally made digitalisation of Formminnesregistret. The Swedish National Heritage Board are working on digitizing the National Ancient Remains register, (Formminnesregistret) into FMIS, a GIS database with all locations and information on a national scale. This is not yet available for Gotland.
2.4 Model Areas

Data for three areas were prepared, each covering 100 km². One area for developing the model, called Training Area. One area for verification of the model, called Verification Area and, finally, one area were the actual predictions would be made, called Prediction Area. The Verification Area was used since the best validation method is Double validation (Independent Data Procedures), which means to use a totally independent dataset to test the performance of the model (Rose & Altschul 1988). The Prediction area is the central agriculture district were hardly any Iron Age settlements are found, but many are believed to have existed.

2.4.1 Training Area
As Training Area, for model development, a 10 x 10 km part of the island was chosen. This area is mostly within the parishes of Kräklingbo and Anga, which are known for holding many remains of Iron Age settlements. The island of Gotland has a very smooth topography, with the highest parts only 82 meters above sea level. The Training Area has a gentle slope towards the sea, with the highest part only 38 meters above the sea in the Southwest part. In Sweden there have been at least two major surveys for archaeological remains covering most part of the country. On Gotland it was done in the 1930’s and the 1970’s. These are very thoroughly done by archaeologists. The remains are obtrusive in their character, and quite easy to spot. In addition to this Carlsson (1979) has surveyed the training and Verification Areas for his thesis in the 70’s, so it can be assumed that very few of the remaining Iron Age foundations are missed.

2.4.2 Verification Area
The Verification Area (10 x 10 km) consists mainly of the parish of Buttle. The area is also known for many remaining Iron Age houses. It is also a well-documented area, which is surveyed several times, and was also a part of Carlsson’s (1979) area of investigation for his thesis. There are 66 sites containing Iron Age houses in the Ancient Remains Database for the area. The area is more forested than the Training Area and has a higher degree of moraine soil covering 56% of the area.

2.4.3 Prediction Area
The objective of this study is two predict areas in the main agricultural districts, so the Prediction area (10 x 10 km) is in the centre of the island, mainly in the parishes of Roma and Barlingbo. The area consists of 57% cultivated land and have sixteen recorded sites with Iron Age houses. The area is very flat, between 25 and 71 meters with only 1% of the area reaching over 55 meters (North-eastern corner).

2.5 Independent Variables
In most predictive models, information derived from DEM’s is an essential part of the variables. Elevation, slope, aspect and other calculated measurements are the basic components in many models (e.g. Stančič 1999; Carmichael 1990; Della Bona 1996, Warren 1990).

Because of the scale and the quality of the presently available DEM’s in Sweden they are not meaningful to use on Gotland in large-scale studies like this. This is due o the flat topography of the island. The Iron Age farmers, and this is also true for most settlers, used higher and dryer ground to build their houses on, but these variations are often very slight, only 1 or 2 meters and consists of very small parts of non-productive land, that don’t show up in the DEM’s. Several tests were made to see if any information could be
used from the DEM, with poor results. This is a problem shared with many others. In Warrens (1990) predictions in Shawnee area in southern Illinois the main cause of the poor result is due to the quality of the Digital Elevation Model (DEM). The DEM was used to derive nine of the variables used. It was derived from digitised contour lines, which were interpolated with a simple algorithm. The DEM produced was terraced (Warren 1990 b). This is a common problem, when simple algorithms are used to produce DEM’s from contour lines (Carrara 1997). Carmichael also identifies the problems with a to rough DEM and to course scale in his study in north-central Montana (Carmichael 1990).

The DEM was only used to derive the Iron Age shoreline. Since Gotland was pressed down by the Pleistocene glacial, and is still rising from the sea, the shoreline of the Iron Age is now up on land. The Iron Age shoreline was estimated to 5 m above the present one in the training area. The verification area and prediction area are not affected, since they are inland areas.

There has been noticed a close correlation between the Iron Age settlements and moraine soils and the meadows in use during historical times. In a study by Carlsson (1979) all, but three parishes reveals that at least 50 % of remaining Iron Age houses are found in 18th century meadows. The meadows are situated in the more wet parts of the infields, and can indicate that animal production was the key production for the Iron Age farmers, since the meadows were used for fodder production. A study made by Arrhenius show that up to 80 % of all Iron Age houses are situated on Moraine soils. It is also noted that the houses are built on, or near parts, of non-productive land, most often the bare rock, which is limestone, the predominant bed rock on Gotland (Carlsson 1979).

<table>
<thead>
<tr>
<th>Iron Age sites</th>
<th>Fields</th>
<th>Meadows</th>
<th>Moraine</th>
<th>18th settlements</th>
<th>Limestone</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of sites</td>
<td>52</td>
<td>52</td>
<td>52</td>
<td>52</td>
<td>52</td>
</tr>
<tr>
<td>Number of sites within 100 m</td>
<td>29</td>
<td>41</td>
<td>39</td>
<td>3</td>
<td>30</td>
</tr>
<tr>
<td>% Of total</td>
<td>56%</td>
<td>79%</td>
<td>75%</td>
<td>6%</td>
<td>58%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Control group sites</th>
<th>Fields</th>
<th>Meadows</th>
<th>Moraine</th>
<th>18th settlements</th>
<th>Limestone</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of sites</td>
<td>58</td>
<td>58</td>
<td>58</td>
<td>58</td>
<td>58</td>
</tr>
<tr>
<td>Number of sites within 100 m</td>
<td>20</td>
<td>19</td>
<td>37</td>
<td>1</td>
<td>16</td>
</tr>
<tr>
<td>% Of total</td>
<td>34%</td>
<td>33%</td>
<td>64%</td>
<td>2%</td>
<td>28%</td>
</tr>
</tbody>
</table>

Table 1 How the sites are distributed within ca.100 meters from the different features. The 100-meter buffer was chosen to compensate for geometrical errors in the different datasets, and can be regarded as possibly located in the same land parcel in reality. The data sets are derived from many different sources and times, using different techniques, so the errors can be relatively large.

The locations of the Iron Age settlements are found in the Ancient remains database kept by the County museum, Gotlands Fornsal. A query was made to extract all records for all “husgrunder” (Stone house foundations). Together with Prof. Dan Carlsson, who has visited nearly all of the house foundations in the training and verification areas in the field, the record set was studied and all the non Iron Age foundations were deleted. He
also decided that the point’s coordinates were representative of the locations of the Iron Age farmsteads. Two records that were situated at the border of the Model area were also deleted, since it was discovered that lay very close to 18th century farms outside the model area, but far from any inside. All other Iron Age settlements had their closest 18th century settlement, meadow and field inside the Model area. All in all 53 points containing Iron Age settlements were used in the model.

The geological soil map presently available for Gotland in digital form is at the scale of 1 to 1 million. The new 1:50 000 soil map is due in digital format in 2004. The scale of 1 to 1 million is not usable in this study, so the old paper soil maps from 1920’s was used. The soil maps for the different areas were transformed and geocoded into the modern Swedish coordinate system, RT 90 and then vectorized manually.

In Sweden, and especially on Gotland, the cadastral maps from 17th century onwards make an excellent source for historical studies. For Gotland we normally have five different time horizons for every farm, spanning from late 1600’s to 1970’s. These maps are suitable for archaeological and land use surveys and analysis. In the 1980’s geographers at Stockholm University digitalized in vector format the land use and the location of the early 18th century farmsteads for the entire island. The vectorisation was based on a complete set of more then three hundred cadastral maps, between 1692 and 1704, covering the whole island. Some parishes were mapped in the 1730-40, but the absolute bulk is from the turn of the century 1700.

The use of the old cadastral maps can be described as a simple and quick paleogeographic method. In the regional predictive map of Kromme Rijn Area, Netherlands, a more complex model of this type was used. Since the landscape changes in that area are greater, a more complex model must be built (Deeben et al 1997).

Another natural feature that might have been important for a prediction model is streams and other fresh water resources. One big problem with streams on Gotland is the heavy drainage by ditches of nearly all bogs and other major wetlands on the island that was carried out in the late 19th and early 20th century. This has altered the streams natural flows, especially in the main agricultural districts. A test was done to produce a stream model from the DEM with the Run-off module in Idrisi32. The result was then compared to the present streams in the Training Area, which is not so heavily affected by the drainage. The result was not good enough to be usable (fig 2). The streams could also be derived from the old cadastral maps, but this was not possible to do within the framework of this work, because it meant to do the entire chain with scanning, rectifying and vectorizing the maps.
Streams generated from the 50 m DEM with the module Run-off in Idrisi
Streams from the modern Economic vector map (Scale 1:20 000)

Fig 2. Comparison between the real stream network and the stream network produced by the Run-off module in Idrisi32. In the southwest corner of the map, the result of the drainage by ditches is visible as straight ditches.

All the variables were reformatted into raster layers of 25 meter resolution covering the Training Area. Distances were calculated from all features. It is the distances to the feature that is the variable, not the feature itself. The same was done with the Verification and Prediction areas.

The variables chosen for the model was:

1. Distance to Moraine soils
2. Distance to 18th century settlements
3. Distance to 18th century meadows
4. Distance to 18th century fields
5. Distance to limestone
6. Bogs and wetlands from the 18th century maps.

For the Training Area a random set of 59 control points was also generated in the GIS and all variables for these points were extracted together with the variables at the sites.
### Table 2. Basic statistics of the site locations (1) and the control group (0). The values refer to distances to the features.

The variables were tested with a T-test in SPSS looking for significant differences between the variables at Iron Age settlements and the control points. The results of the T-Test showed significant differences between the two groups concerning meadows, fields, limestone and 18th century sites. For moraine there was no statistical difference. The none-significance for the moraine can be explained by the fact that 40.6% of the model area consists of moraine soils that are evenly distributed over the model area. Since moraine is a vital variable it was kept in the model.

---

1 T Test procedure compares means for two groups of cases. The mean values for the two groups are displayed in the Group Statistics table. If the significance value for the Levene test is high (typically greater than 0.05), then use the results that assume equal variances for both groups. If the significance value for the Levene test is low, then use the results that do no assume equal variances for both groups. A low significance value for the t test (typically less than 0.05) indicates that there is a significant difference between the two group means. If the confidence interval for the mean difference does not contain zero, this also indicates that the difference is significant. If the significance value is high and the confidence interval for the mean difference contains zero, then you cannot conclude that there is a significant difference between the two group means (SPSS manual).
### Table 3. Results from the T-test. The low values in the sig. column of the Levene's test suggest that significant differences between the control group and sites are present for meadows, fields and 1700’s sites. Also the T-test’s 95% confidence interval and sig column indicates the same thing. No one of the three mentioned variables contains zero in the interval. The none significance for moraine can be explained by the fact that 40.6% of the Training Area consists of moraine soils that are evenly distributed over the model area

### 2.6 Model Development

The program used for the Logistic Regression Analysis was XLSTAT 6.0. It was preferred over the LOGITREG module in Idrisi32, since it reported more adequate statistics about the result. Idrisi32 only produced a T-test for the variables and an F-test for the model fit, and these are not the preferred tests. Chi$^2$ test that is the correct method for verifying the variables and overall fit of the model. All the variables were taken into the first LRA model. The result is seen in table 4 below. Only two variables, Distances to meadows and distances to 18$^{th}$ century sites are significant (a lower value then 0.05 in the Pr.> Chi2 column).

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimates</th>
<th>Std. deviation</th>
<th>Chi2</th>
<th>Pr. &gt; Chi2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>2.510599</td>
<td>0.598230</td>
<td>17.612367</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>Fields</td>
<td>0.001851</td>
<td>0.001482</td>
<td>1.558948</td>
<td>0.211819</td>
</tr>
<tr>
<td>Meadows</td>
<td>-0.005749</td>
<td>0.001502</td>
<td>14.656879</td>
<td>0.000129</td>
</tr>
<tr>
<td>Moraine</td>
<td>-0.000225</td>
<td>0.001769</td>
<td>0.016135</td>
<td>0.898921</td>
</tr>
<tr>
<td>18th sites</td>
<td>-0.001741</td>
<td>0.000870</td>
<td>4.003913</td>
<td>0.045395</td>
</tr>
<tr>
<td>Limestone</td>
<td>-0.002851</td>
<td>0.001591</td>
<td>3.210744</td>
<td>0.073156</td>
</tr>
</tbody>
</table>

Table 4. Estimates of the parameters of the first model (maximum likelihood)
All the non-significant variables except moraine were dropped. Moraine was kept due to its great importance (see above). The fact that it gets such a low significance is due to the large areas of moraine. The influence of the moraine variable in the model is still not much, due to its low estimate value. The final model is shown in table 5.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimates</th>
<th>Std. deviation</th>
<th>Chi2</th>
<th>Pr. &gt; Chi2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>2.125308</td>
<td>0.510601</td>
<td>17.325317</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>Meadows</td>
<td>-0.005348</td>
<td>0.001463</td>
<td>13.369384</td>
<td>0.000256</td>
</tr>
<tr>
<td>Moraine</td>
<td>0.000815</td>
<td>0.001618</td>
<td>0.253995</td>
<td>0.614276</td>
</tr>
<tr>
<td>18th sites</td>
<td>-0.001617</td>
<td>0.000634</td>
<td>6.496928</td>
<td>0.010806</td>
</tr>
</tbody>
</table>

Table 5. Estimates of the parameters of the final model (maximum likelihood)

In table 6 we see the estimated fit of the model. The table shows that the significances is quite poor, but compared to the independent model, which is a model generated by the program with only the intercept, it shows that the model is significant and the independent variables bring important information to the model. The $R^2$ values cannot be read like in an ordinary Linear Regression model, it is only a pseudo $R^2$. The McFadden $R^2$ is closer, but can still only give a hint of the fit. A mid value is normally considered very good (Kvamme 1988 b).

The variables and the model may not meet all statistical demands. The variables are not completely independent from each other. The 18th settlements have a correlation to the meadows. Normally the meadows are located in the vicinity of the house-lots. They’re are exceptions to this rule in the case of farmsteads are merged with others and the abandoned farms fields are kept in production. Since the Moraine is such a large part of the area and is the preferred soil type for meadows, it is also correlated to meadows. These statistics can only give a clue to how the model will perform; the most important test will be how it predicts the Verification Area. To quote Kvamme (1988):

“When I asked them [professional statistician] about the role of statistical theory in model development, they suggested that I worry less about theory and more about how well the model works in practice”

<table>
<thead>
<tr>
<th>Obs</th>
<th>Chi²</th>
<th>df (Chi²)</th>
<th>Pr. &gt; Chi²</th>
<th>L.R. Chi²</th>
<th>df (L.R. Chi²)</th>
<th>Pr. &gt; L.R. Chi²</th>
<th>R²</th>
<th>R² (McFadden)</th>
</tr>
</thead>
<tbody>
<tr>
<td>110</td>
<td>119.8</td>
<td>106</td>
<td>0.170</td>
<td>55.410</td>
<td>3</td>
<td>&lt; 0.0001</td>
<td>0.438</td>
<td>0.364142</td>
</tr>
</tbody>
</table>

Table 6 Evaluating the goodness of fit of the final model

The Logistic Regression formula derived and used in the analysis:

$$\text{Prob.} = \frac{1}{1 + \exp(-(2.125308 -0.005347 \times \text{Meadows} + 0.000815 \times \text{Moraine} -0.001616 \times \text{18th sites}))}$$
2.7 Results

The formula was entered into the Image Calculator in Idrisi32 and calculated for the Training Area. The resulting image was then overlaid with the bogs and the sea level for the Iron Age, to cut away impossible areas. A cut point at 0.5 was chosen. This means that the areas with a probability of over 0.5 were considered to be the model area, which is common practice. The gain was then calculated according Kvammes formula (Kvamme 1988):

\[
\text{Gain} = 1 - \left( \frac{\text{percentage of total area covered by model}}{\text{percentage of total sites within model area}} \right)
\]

The area covered by the model was 23% and 88% of the sites were predicted within that area. This gives a gain of 74%, which must be considered good. (If the gain is calculated before the overlay operations, the model area is 28% at a 0.5 cut-point and the gain is 68%).

The same operations were then made with the Verification Area, the formula was calculated in Idrisi32 and the overlay with the bogs were performed. The Verification Area contains no sea shoreline and no overlay with the sea was needed. The performance cannot be expected to as high for the Verification Area, as for the Training Area. This is because the optimisation of the models coefficients (estimates) uses all oddities of a particular data set. The data set used for model development will almost always perform better then any other data set (Rose & Altschul 1988).

In the Verification Area the model cover 26% of the area and it predicted 74% of the sites after the bogs were removed, at the 0.5 cut-point. The gain was 65%, which is less then for the Training Area, but must be considered good. The area, before the bogs were removed, covered by the model was 28%. The gain was 62%.

Finally, the central agricultural district will be predicted, in the Prediction area. The procedure is the same as with the two other areas. The Prediction Area is not connected to the sea. The result of the Predictions is seen in Fig 3. 53% of the area fell within the model, at the 0.5 cut-point. All but two of the known locations of Iron Age houses were predicted right.
Fig 3 The result of the predictions in the Prediction Area. This map is in the national Swedish grid, RT 90, and could easily be combined with the modern economic or Topographic map, to create a survey map in larger scale.
2.8 Concluding discussion and future work

Considering the very few independent variables used the results from this exercise were better than expected. One of the most commonly used sources for Predictive Modelling was not used, a DEM. Due to the topography of Gotland and the present quality of DEM’s they are unusable in a large-scale model like this. With this work we have shown that the cadastral maps can be used as sources of information in a GIS for predicting events and time periods, long before the production of the map. In many areas of Gotland, the changes that created the present landscape are quite late in time and occurred during the late 19th and 20th centuries.

If we refer to the discussion above, about types of approaches, this study can be described more of an inductive then deductive study. The variables used are more of observed then explanatory. But in some sense you can also see the variables used as explanatory, since they stress the long continuity of the farming on Gotland. The same areas used 18-1 500 years ago were still in use 1 000 years later. Whether or not it is for the exact same purpose or not is debated. But the same areas were settled and used for agriculture, only the house lots had moved. (Carlsson 1979). It was the most fertile soils that were first used. Carlsson (1979) can see a direct link between fertile soils and the earliest agricultural traces.

As expected there were a number of sites, outliers, which could not be predicted. These are what Altschul (1990) calls ‘red-flags’ which has been discussed earlier. They are situated a bit from the central areas in each parish. These are mainly the ones that were abandoned during the agricultural crisis of the 6th century. Some of them lived on into the Middle Ages, and even further (Carlsson 1979). Most of these were not predicted and they need further analysis, maybe a model of their own. According to Della Bona’s (1994) schema for model development (Table 1), we have in this paper only done the two first stages. To be able to predict the outliers, if ever possible, means that the Tertiary Stage in his schema has to be implemented. This was not possible within the timeframe and budget for this study. Questions that need to be addressed in any analysis of the outliers are for example: What is the weakness of the abandoned sites? Why the expansion of the 17th –18th century choose other areas and not the ones abandoned during the Iron Age? What are the differences between the sites and the agricultural economy and technique between the two Eras? Many of these questions are already answered by Carlsson and others (Carlsson 1979), but how to implement the answers in a GIS and a Predictive model is not yet done.

In the 30’s and 40’s Professor Arrhenius did a lot of phosphate mapping on Gotland. This was done in order to find the best soils for sugar beets. He soon discovered that it also was a map over areas of Human activities, especially settlements. These maps are available in paper copies and could be incorporated into the model. One problem is that you cannot date the phosphate areas, but together with the other variables, we think that the model could be refined with this data. Another source could be the Cadastral maps produced under the Land Reform Act of 1827 (Laga skifte). In these maps the absolute
majority of the island is mapped in the scale of 1:4 000 over a period of more than 100 years. Many consider these maps the best ever produced in Sweden. In them, all pieces of land are classified and its fertility and other characteristics are noted in accompanying descriptions. To transfer them into digital databases is however a huge task.

In the future, when digital data for the entire island is available, a prediction model for the entire island will be done. This kind of location model can be used for a number of studies and purposes. One task would be to try to calculate the number of disappeared stone house foundations, and getting a better estimate, the ones existing (see discussion above). Since we on Gotland also have a very good picture of the settlement distribution around 1700 this data can be analysed versus a location model of the older Iron Age. For example can regional division be analysed and the continuity of overall settlement structure. Also data from other time periods, like Viking Age data based on the work of Östergren (1989) can be compared with the model. The options are many, and with the rich archaeological remains of Gotland, this kinds of studies are hard to make in any other part of Gotland.

The method chosen, Logistic Regression Analysis, is probably not the only method suited for this kind of problem. In studying the result map a simpler overlay method may have produced quite a good result. There are two important variables: the distances from 18th century meadows and settlements, which quite easily could be handled in an overlay analysis. But the outcome of a LRA as a continuous surface with probabilities is a major advantage over overlay methods, so it was chosen. However we think that there are other data, which could be included in the model, but wasn’t tried due to time limits.

As mentioned above, the variables do not meet all statistical requirements for some of the statistics used. But these requirements are probably seldom met in these kinds of models. There have been tests suggesting that models can perform well without being perfect in their underlying statistical assumptions (Kvamme 1990). This model performed well, without being perfect in a statistical sense.
References


Dalla Bona, Luke


Kvamme, Kenneth L.


Warren, Robert E.


