Journal of Archaeological Science 92 (2018) 28–36

Contents lists available at ScienceDirect

Journal of Archaeological Science

journal homepage: http://www.elsevier.com/locate/jas

Predictive modeling for archaeological site locations: Comparing logistic regression and maximal entropy in north Israel and north-east China

Ido Wachtel [a,*], Royi Zidon [b], Shimon Garti [c], Gideon Shelach-Lavi [d]

[a] The Institute of Archaeology, The Hebrew University of Jerusalem, Mt. Scopus, Jerusalem, 91905, Israel
[b] Department of Evolution, Systematics, and Ecology, The Alexander Silberman Institute of Life Sciences, The Hebrew University of Jerusalem, Israel
[c] Einstein Institute of Mathematics, The Hebrew University of Jerusalem, Israel
[d] Department of Asian Studies, The Hebrew University of Jerusalem, Israel

ARTICLE INFO

Article history:
Received 16 July 2017
Received in revised form
10 January 2018
Accepted 1 February 2018

Keywords:
Northern Israel
Northeast China
Predictive modeling
Logistic regression
Maximal entropy (MaxEnt)

ABSTRACT

Archaeological predictive modeling is a tool that helps assess the likelihood of archaeological sites being present at different locations in the landscape. Such models are used for research purposes, as an analytical tool to better explain settlement patterns and past human behavior. They are also an important tool for the preservation of archaeological sites, as they can help planners avoid areas where sites are likely to exist. In this study we compare two methods of predictive modeling for archaeological site locations using two independent case studies. The more commonly used method of *logistic regression* is compared with a newer method of maximal entropy (*MaxEnt*). We examine the effectiveness of both models on two independent datasets collected from the Upper Galilee (northern Israel) and the Fuxin area (northeast China). While both methods have proven useful, in both cases the MaxEnt models produced much better results, which were much more efficient, than those of the logistic regression.

1. Introduction

Archaeological predictive modeling is a tool that helps to assess the likelihood of archaeological sites being present at different locations in the landscape (Warren and Asch, 2000: 6; Mehrer and Wescott, 2006; Kvamme, 2006). The use of such models is well-established in archaeological research since the 1980s, mostly in the field of contract archaeology and in efforts to prevent the destruction of sites (Wheatley and Gillings, 2002: 148). With the rapid increase in computation capabilities and the improvement of modeling methods it holds greater unrealized potential as a powerful tool for both contract archaeology and analytical research.

The basic assumption in archaeological predictive models is that the location of ancient sites is not random, but rather reflects human choices, and is influenced by the natural conditions and the availability of natural resources. Another premise is that the environmental variables of the ancient sites are still present in the landscape, and can be measured and quantified using modern maps, satellite images and other geographical sources (Warren and Asch, 2000:7; Kvamme, 2006: 4). The predictive model performs a quantitative analysis of the environmental factors associated with a sample of known sites, and projects those factors onto areas where sites are unknown as a probability surface. Those surfaces can be represented graphically as a colored or binary maps of high and low values for the probability of sites being found (Kvamme, 1988, 1990: 261). While such models are most commonly used to identify areas where sites are likely to be found and thus protect them from destruction by construction projects, they can also be used for research purposes, as an analytical tool to better explain settlement patterns and past human behavior (Judge and Sebastian, 1988; Kvamme, 1990, 1992, 2006; Warren and Asch, 2000; Hudak et al., 2000; Wescott and Brandon, 2000; Wheatley and Gillings, 2002).

In the present article we attempt to contribute to this field by comparing two methods of predictive location modeling applied to two independent sets of data from two areas – North Israel and Northeast China. In both areas, we use the same approach: the more traditional method of logistic regression is compared with the newer method of maximal entropy (MaxEnt). These case studies were chosen because they represent the two scenarios where predictive models are most likely to be used: (1) When some, but

* Corresponding author.
E-mail address: ido.wachtel@mail.huji.ac.il (I. Wachtel).

https://doi.org/10.1016/j.jas.2018.02.001
0305-4403/

probably not all, archaeological sites are known from a relatively large region and we wish to predict where yet unknown sites may be found within this region. (2) When most of the archaeological sites are known from a relatively small area and we wish to project this knowledge and predict where sites may be found in adjacent regions where no archaeological work has been done. The results of our experiments clearly demonstrate that in both type of cases the predictive power of MaxEnt and its potential for use in archaeological research is much greater than that of traditional logistic regression.

2. Model theory and background

The predominant statistical technique in constructing archaeological predictive models is *logistic regression*. As it has often been discussed in archaeological publications (Kvamme, 1990; Warren, 1990; Wescott and Brandon, 2000) it is sufficient for present purposes to say that this kind of probability model is suitable where the dependent variable is binary (yes/no). The results of the model express the probability of an 'event' (such as archaeological site) in terms of probabilistic value between 0 and 1. The environmental parameters are defined as independent variables and the model examines the relationship between the independent variables and the dependent variables and calculates the probability value of the dependent variable at all the points of the map (for more details see appendix).

In principle, logistic regression requires presence-absence data. But, in the majority of cases in archaeological predicting models, our data is positive, i.e., we can tell, based on preliminary archaeological surveys, where archaeological sites do exist. We cannot, however, determine that the absence of data indicates the absence of sites. We cope with this problem by randomly choosing points on the map and declaring them “non-sites.” The rationale underlying this procedure is that sites are rare “events” in the landscape, and cover a relatively small area, usually less than 1% of the region. Therefore, even if we were to arbitrarily choose 100 random points, almost all of them would fall on a “real” non-site location (Kvamme, 1992).

Maximum Entropy Modeling has been used in urban geography and archaeological research since the 1960s (Wilson, 1970; Evans and Gould, 1982; Bevan and Wilson, 2013; Davies et al., 2014; Altaweel, 2015; Howey et al., 2016). However, in most cases, MaxEnt (or other methods which use similar principles) has been used for the analysis of dynamic processes and spatial interactions. In the current paper we suggest employing MaxEnt as a tool for predicting site locations (Phillips et al., 2004, 2006; 2018; Elith et al., 2011).

The MaxEnt method is based on two principles:

(1) The expectancies comparison principle.
(2) The maximal entropy principle.

The expectancy of a variable (e.g., elevation or soil type) is parallel to the mathematical concept of an average.

A reasonable first step for finding a good probability map would be to compare the expectancies *of all variables*. Let $\widehat{\pi}$ be the distribution extracted from the observations and let Y be any parameter (e.g., temperature at every point); and let $E_{\widehat{\pi}}(Y)$ be the expectancy of Y according to $\widehat{\pi}$. Let π be the real distribution that we are looking for. We require that:

(i) $E_{\widehat{\pi}}(Y) = E_{\pi}(Y)$ for every parameter Y.

The collection of all π which satisfy (i) is typically large. In non-trivial cases, the number of these distributions is infinite. How are we to choose the most appropriate one from this infinite collection of distributions?

Here we invoke the second principle of maximal entropy. In what follows, $\times$ is the free variable which represents the location of a point. Entropy is defined as follows:

(ii) $H(\pi) = -\sum_X \pi(x) ln \pi(x).$

The principle of maximal entropy instructs us to choose from among all appropriate distributions the one with the maximal value of entropy.

The mathematical concept of entropy reflects disorder, or freedom. The basic assumption is that natural phenomena tend to disorder, unless a directed effort is made in the opposite direction. In the context of archaeological sites, people choose a place to establish their settlement in a directed way and its location is not random. However, apart from that directed effort we expect freedom or randomness. By choosing a distribution whose expectancies coincide with the expectancies of the observations, we satisfy the directed components. By isolating the distribution with maximal entropy we ensure that all other considerations are maintained random and there is no other influence on our map.

Generally speaking, one has to proceed through the following stages in implementing the principals of MaxEnt:

1. Find the distribution $\widehat{\pi}$ of the observations data set (archaeological sites).
2. Calculate the expectancies of $\widehat{\pi}$ with respect to all the relevant parameters (predictors).
3. Identify the set A of all distributions with the same expectancies as $\widehat{\pi}$.
4. From among the members of set A, identify the distribution with maximal entropy.

Let t be the total number of points in the map. For each point we set

(iii) $\widehat{\pi}(x) = \begin{cases} 1/t & \text{if there is an observed site at x} \\ 0 & \text{otherwise} \end{cases}$

Notice that the sum of all values of $\widehat{\pi}$ is exactly one, indicating that this is indeed a probability vector.

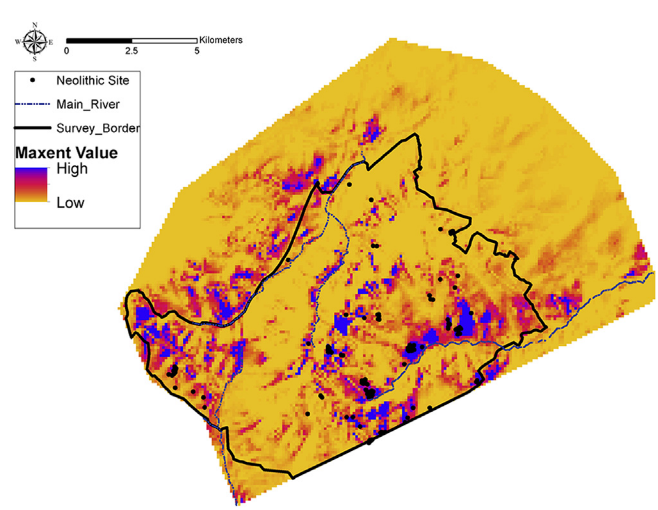

Fig. 1. Color map of a MaxEnt model showing predicted locations for Neolithic sites across the Fuxin region, Northeast China. Pink and blue pixels represent areas of high probability for site presence. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)