FORCASTING STATISTICAL MODELS OF ARCHAEOLOGICAL SITE LOCATION

1. INTRODUCTION

Forecasting statistical models are becoming increasingly important in archaeological research (KOHLER, PARKER 1986). One of the reasons of this popularity is that archaeological sites tend to present themselves in particular environments so that forecasting models can help in identifying areas where the probability is higher based on previously collected statistical information. In the present paper we will consider a class of statistical models designed to produce maps of the probability for archaeological site location (henceforth ASL) which incorporate both deductive and inductive considerations.

Forecasting models for the probability of ASL can be classified into two classes by distinguishing between models on a continuous space and models on a discrete space. The output produced in the two cases is displayed in Fig. 1. In the first instance the models produce a probability surface for ASL (Fig. 1a). In the second instance the space is discretized by superimposing a grid of contiguous quadrats and the output is an array of probability values (Fig. 1b).

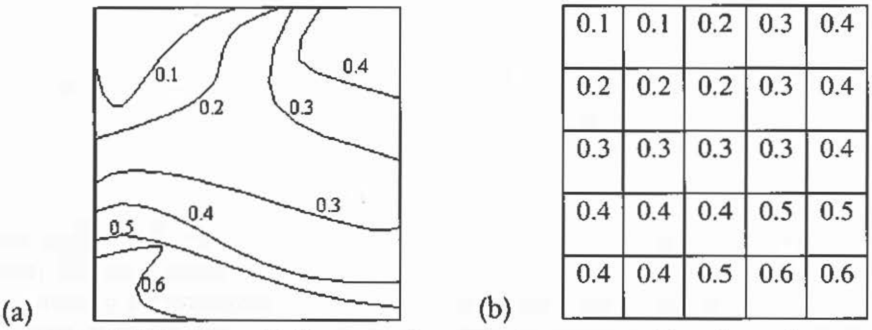
In the present paper we will refer about the class of statistical models on a discrete space. The approach based on a continuous space have been exploited elsewhere (for instance by BENEDETTI, ESPA 1995), but is not considered here.

The models described can be of help in practical circumstances exploiting the potentialities of data derived from satellite and aerial photographs and can be easily implemented within the context of a GIS (ARROYO-BISHOP 1995).

The paper is organised as follows. Section 2 is devoted to discuss one of the currently most popular approach to forecasting modelling in archaeology that is the "integrated strategy" of KWAMME (1983) and WARREN (1990). Starting from the weakness of this approach in Sections 3 and 4 we will discuss possible extensions to correct the procedures using contextual information, we propose a new methodology and we discuss the relative advantages against the current practice. Finally Section 5 is devoted to some conclusions and to outline the research agenda in the field.

2. THE INTEGRATED APPROACH: LOGISTIC MODELS

One of the most popular approach to forecasting modelling in archaeology is the "integrated strategy" developed by KWAMME (1983) and employed by several researchers like WARREN (1990). Kwamme strategy exploits the potentiality offered by a GIS to create and process large data sets through the logistic



0.1	0.1	0.2	0.3	0.4
0.2	0.2	0.2	0.3	0.4
0.3	0.3	0.3	0.3	0.4
0.4	0.4	0.4	0.5	0.5
0.4	0.4	0.5	0.6	0.6

Fig. 1 – Probability maps of ASL expressed on a continuous space (a), and on a discrete space (b).

regression, a flexible statistical tools which allows to forecast binary variables.

Suppose we are analysing a study area where a number of ASL have been identified by means of a field survey. The study area is discretized into *M* contiguous quadrat cells. Suppose that in *N* of such cells (*N*<*M*) the site survey was able to assess the presence or absence of an archaeological site. On this basis we define a variable Y_i such that:

$$Y_i = \begin{cases} 1 \text{ if cell } i \text{ contains an archaeological site} \\ 0 \text{ otherwise} \end{cases} \qquad i = 1,\dots, N$$

An example is reported in Fig. 2.

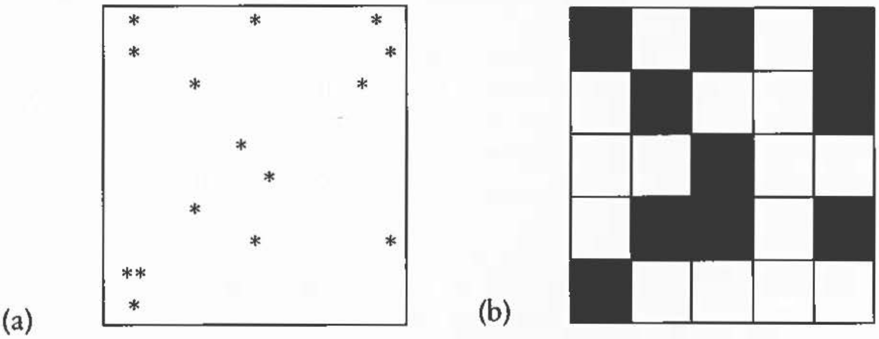


Fig. 2 – Continuous distribution of ASL in a study area (a) and its discretized version on a 5-by-5 quadrat grid (b).

Suppose further that for all *M* cells it is available a set of auxiliary information about k independent variables $\mathbf{X}_i \equiv (X_{1i}, X_{2i}, \dots\dots, X_{ki})$ i = 1,....., *M*. The independent variables could be related to the nature of the soil (slope, erosion, exposure, etc.), to the topography (DTM), the hydrology (rivers, channels, lakes, etc.), the topology (nearness to streets, to communication routes, etc.), or to other variables derived from existing cartography, aerial or satellite imagery, survey samples and other sources (CARLA' *et. al.*, 1995).

In such a situation we can define on the basis of the N observations a model in which the probability of finding an ASL, say θ_i = Prob{Y_i = 1}, is a function of the vector of predictors X and of a vector γ of parameters, that is:

$$[1] \quad \theta = f(X, \gamma)$$

In particular the linear logistic regression (COX, SNELL 1989) specifies the relationship [1] as

$$[2] \quad \Pr ob\{Y_i = 1\} = \theta_i = \frac{\exp[x'\gamma]}{1 + \exp[x'\gamma]}$$

and similarly

$$[3] \quad \Pr ob\{Y_i = 0\} = \theta_i = \frac{1}{1 + \exp[x'\gamma]}$$

or, in general,

$$[4] \quad \Pr ob\{Y_i = \gamma_i\} = \frac{\gamma_i \exp[x'\gamma]}{1 + \exp[x'\gamma]}$$

Model [2] and [3] can be estimated via a maximum likelihood procedure by choosing a subset n (n<N) of the N selected cells (called *training sites*), and cross-validated by contrasting the results with the (N-n) observed sites. Finally the model can be employed to forecast the probability of ASL on the (M-N) unobserved cells. The output is a probability map of the kind displayed in Fig. 1b.

However the use of the logistic regression in such a context is statistically incorrect since the model assumes a spatial independence in the Y_i's (COX, SNELL 1989), an hypothesis which is patently violated in all geographical studies where (as stated in Tobler's – first law of geography –, TOBLER 1970) «everything is related to everything else, but near things are more related than distant things». As a role archaeological sites tend to cluster in space and this fact results into higher probability of sites in the neighbourhood of existing sites. As a consequence closeness to other archaeological sites modifies (generally increases) our expectation of other sites. On the other hand it is obvious that neglecting such a contextual information is statistically inefficient and is doomed to produce unreliable estimates and forecasts (see CRESSIE 1992; ARBIA 1993).

3. IMPROVING THE INTEGRATED APPROACH: THE AUTO-LOGISTIC MODEL

The formal way of incorporating the notion of contextual information into binary response variable models has been introduced in the statistical literature by BESAG (1974).

Define for each cell *i* constituting a study area (*i* = 1, 2, ...M), the set N(i) which represents the set of neighbouring cells. For instance one could assume a neighbouring scheme on horizontal and vertical direction (or "rook's neighbourhood" from the chess rook's move) as the one depicted in Fig. 3:

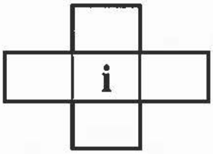


Fig. 3 – Neighbouring cells to the *i*-th cell with a horizontal/vertical criterion.

Now we can model the probability of finding a cell containing an ASL conditional upon the observations of the remaining cells as a function of only the neighbouring cells N(i), and of a set of parameters β

$\theta_i = \text{Prob}\ \{Y_i = 1 \mid Y_j = 1\ j \neq i\} = f\ (Y_j\ , j \in N(i), b)$

Besag's "auto-logistic" model is expressed as:

$$[5] \quad \theta_i = P\left(Y_i = y_i \mid y_j, j \in N(i)\right) = \exp\left[y_i\left(\beta_{0i} + \sum_{j \in N(i)} \beta_{ij} y_j\right)\right] \Big/ \left[1 + \exp\left(\beta_{0i} + \sum_{j \in N(i)} \beta_{ij} y_j\right)\right]$$

where y_i =[0,1], N(i) represents the sets of cells close to the i-th pixel of the image and the β_{0i} and the b_{ij}'s are parameters to be estimated. In particular β_{0i} is a location parameter which can incorporates the influence of the independent variables selected from theory, and the remaining β's control for spatial interaction, that is the intensity of the influence of local context on archaeological site expectations (Arbia 1993).

Model [5] incorporates the notion of Markov dependency in spatial processes and is also known in statistical physics as *Ising's law* (see Ising 1925 or Arbia 1993 for a review). The spatial interaction parameters β_{ij}'s can often be decomposed as

$$[6] \quad \beta_{ij} = \beta\, w_{ij}$$

with w_{ij} such that:

$$[7] \quad w_{ij} = \begin{matrix} 1 \text{ if } j \in N(i) \\ 0 \text{ otherwise} \end{matrix}$$

The parameter β embodies the degree of contextual information accounted for in the model. Consider the following example. Suppose that in Formula [5] we have β_0 = 1 each i, and $\beta_{ij} = \beta\, w_{ij}$, and consider the following distribution of presence/absence of ASL reported in Fig. 4.

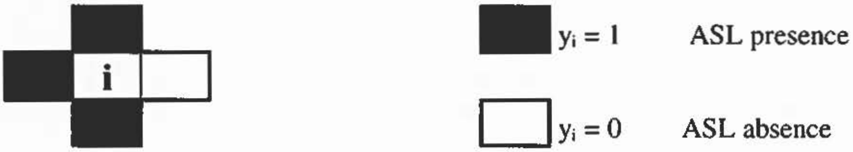


Fig. 4 – Distribution of ASL Presence/Absence in the neighbourhood of site *i*.

In this case Expression [5] becomes:

$$[8] \quad \theta_i = \Pr ob(Y_i = 1) = \frac{\exp(1+3\beta)}{1+\exp(1+3\beta)}$$

If there is no contextual information captured by the model we have that β = 0 and Formula [8] becomes:

$$\theta_i = \Pr ob\ (Y_{.} = 1) = \frac{\exp(1)}{1 + \exp(1)} = 0.73$$

In contrast, if there is a positive expectation of finding ASL close to one another, we have that β > 0. For instance, if β = 1/3 we have:

$$\theta_i = \Pr \mathrm{ob}(Y_i = 1) = \frac{\exp(2)}{1+\exp(2)} = 0.88$$

and the probability is higher than in the case of no contextual information. If we want to account for a higher degree of contextual information, for instance β = 1, we have, instead:

$$\theta_i = \Pr ob(Y_i = 1) = \frac{\exp(4)}{1+\exp(4)} = 0.98$$

which further increases the probability.

The model described above can be estimated via a pseudo-maximum likelihood procedure or, alternatively through the so-called *coding technique* (Besag 1974) by choosing a subset *n* (*n*<*N*) of the *N training sites,* cross-validated by contrasting the results with the (*N*-*n*) observed sites, and employed to forecast the probability of ASL on the (*M*-*N*) unobserved cells.

4. A joint logistic/autologistic approach

The approach presented in the previous section suggests that a gain in the comprehension of ASL can be obtained by considering information collected in the context where each observation is embedded. However, contrasting the logistic model presented in Section 2 which is statistically incorrect, we now have a model which is statistically correct, but does not take into account in a sufficient way of any *a priori* knowledge on ASL derived from auxiliary variables. It seems therefore reasonable to join the two approaches and to introduce a third model that seeks to incorporate both contextual and auxiliary information.

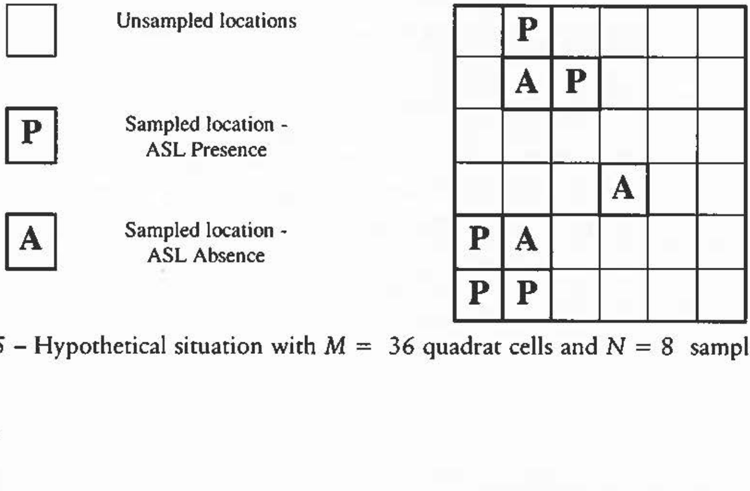
The *log/autolog* model (as we will henceforth refer to this third model) states that the probability of finding an ASL conditional on existing data and on contextual information can be represented as:

$$[9] \quad \theta_i = \text{Prob}\ \{Y_i = 1 \mid y_j\ j \in N(i);\ X\} = \frac{\exp(\beta_0 + \sum_j \beta_{ij} y_j + X'\gamma)}{\left[1 + \exp(\beta_0 + \sum_j \beta_{ij} y_j + X'\gamma)\right]}$$

where β_0 is a constant, the β_{ij}'s are the spatial interaction parameters with an analogous meaning to those introduced in the autologistic model (see Formula [8]), the vector **X** is a vector of independent variables like those employed in model [2] and [3] related to auxiliary information, and γ the associated parametrization.

Model [9] is conceived so as to remove the problems connected with the non independence of the Y's, and simultaneously, to augment the autologistic model with information coming from auxiliary variable related to the nature of the soil. As such it appears ideal under many respects in that it manages to produce more reliable probability maps of ASL than the logistic model used following the "integrated approach", and to incorporate archaeologists' prior knowledge on the study area. However, as we did in the previous sections while presenting the various models, it is fair to underline the drawbacks of this alternative approach.

The main one seems to be that model [9] is formally more difficult to treat than the other two models, and that some of the statistical problems connected with its estimation are still unsolved. The statistical estimation problem can be described as follows. The usual situation we have to deal with is the case in which we have the study area discretized into *M* cells, but observations are available for only *N* of such cells (see Fig. 5). Since observa-


□ Unsampled locations
P Sampled location - ASL Presence
A Sampled location - ASL Absence

	P				
	A	P			
			A		
P	A				
P	P				

Fig. 5 – Hypothetical situation with *M* = 36 quadrat cells and *N* = 8 sampled locations.

tions are not available for all the random variables which constitute the spatial process, we are in the statistical situation termed "incomplete-data problem". Some solution to this problem has been proposed in the literature (see DEMPSTER, LAIRD, RUBIN 1977), and modifications to account for spatial Markov dependency have been attempted (QIAN, TITTERINGTON 1989; RATHBUN, CRESSIE 1994). However the effectiveness in the case in hand needs to be studied and tested in practical cases.

5. CONCLUSIONS AND RESEARCH AGENDA

In the present paper we have criticised the use of the logistic regression for the production of ASL probability maps proposed by KWAMME (1983), an approach known in the archaeological literature as the "integrated strategy". The application of the method is statistically incorrect since in archaeological studies it is violated the hypothesis of independence between cells which is at the basis of the logistic regression model.

To overcome such limitations we have proposed two alternative models. The first one is an autopredictive model in which the probability of ASL is modelled as a function of the observations coming from field surveys in neighbouring zones. This approach accounts for the problem of non-independency of observations, but neglects *a priori* auxiliary information on the archaeological area. The second approach is a more comprehensive one which overcomes the problems of logistic regressions while preserving the role of *a priori* information.

The research agenda in the field presents three major issues. First of all it is necessary to test the autologistic model (Section 3) and the log/autolog model (Section 4) on real data sets. Secondly the results obtained on a set of training sites need to be compared with those obtained by using the standard "integrated approach" based on logistic regression. Finally a purely statistical problem needs to be faced while devising good estimators for the log/autolog model where (as described in Section 4) we face an incomplete-data problem.

GIUSEPPE ARBIA
Department of Quantitative Methods and Economic Theory
University "G. D'Annunzio" - Pescara

GIUSEPPE ESPA
Institute of Statistics and Operational Research
University of Trento

BIBLIOGRAPHY

ARBIA G. 1993, *Recenti sviluppi nella modellistica spaziale*, in S. ZANI (ed.), *Metodi Statistici per le Analisi Territoriali*, Milano, Franco Angeli, 193-217.

ARROYO-BISHOP D. 1995, MEDAIS: *The potential contribution of GIS to the identification,* *study and conservation of the cultural heritage of the Mediterranean basin,* in *Proceedings of the First International Congress on Science and Technology for the Safeguard of Cultural Heritage in the Mediterranean Basin*, Catania-Siracusa, 103.

Benedetti R., Espa G. 1995, *The use of remotely sensed images as auxiliary information for the interpolation of archaeological data,* in *Proceedings of the First International Congress on Science and Technology for the Safeguard of Cultural Heritage in the Mediterranean Basin*, Catania-Siracusa, 115.

Besag J.P. 1974, *Spatial interaction and the statistical analysis of lattice systems* (with discussion), «Journal of the Royal Statistical Society», 36, B, 192-235.

Carla' R., Carrara A., Jacoli M., Alessandro V., Baronti S. 1995, *Analysis of multispectral imagery for archaeological investigation,* in *Proceedings of the First International Congress on Science and Technology for the Safeguard of Cultural Heritage in the Mediterranean Basin*, Catania-Siracusa, 129.

Cox D.R., Snell E.J. 1989, *Analysis of Binary Data*, 2nd ed., London, Chapman & Hall.

Cressie N. 1992, *Statistics for Spatial Data*, New York, John Wiley and Sons.

Dempster A.P., Laird N.M., Rubin D.B. 1977, *Maximum likelihood for incomplete data via the EM algorithm,* «Journal of the Royal Statistical Society», B, 39, 1-22.

Ising E. 1925, *Beitray sur theorie des ferromagnetismus,* «Zeitschrift Physic», V31, 253-258.

Kohler T.A., Parker S.C. 1986, *Predictive models for archaeological resource location*, in M.B. Schiffer (ed.), *Advances in Archaeological Method and Theory*, New York, Academic Press, 9, 397-452.

Kwamme K.L. 1983, *Computer processing techniques for regional modelling of archaeological locations*, «Advances in Computer Archaeology», 1, 26-52.

Qian W., Titterington D.M. 1989, *On the use of Gibbs Markov chain models in the analysis of images based on second-order pairwise interaction distributions,* «Journal of Applied Statistics», 16, 267-281.

Rathbun S.L., Cressie N. 1994, *A space-time survival process for a longleaf pine forest in Southern Georgia,* «Journal of the American Statistical Association», 89, n. 428, 1164-1174.

Tobler W.R. 1970, *A computer movie simulating urban growth in the Detroit Region*, «Economic Geography», suppl. 46, 234-240.

Warren R.E. 1990, *Predictive modelling of archaeological site location: a case of study in the Midwest*, in K.M.S. Allen, S.W. Green, E.B.W. Zubrow (eds.), *Interpreting Space: GIS and Archaeology*, New York, Taylor & Francis, 90-111.

ABSTRACT

Forecasting statistical models are becoming increasingly important in archaeological research. One of the reasons of this popularity is that archaeological sites tend to present themselves in particular environments so that forecasting models can help in identifying areas where the probability is higher based on previously collected statistical information.

In the present paper we consider a class of statistical models designed to produce maps of the probability for archaeological site location (ASL) which incorporate both deductive and inductive considerations. In the discussion we criticise the use of the logistic regression for the production of ASL probability maps, a popular approach known in the archaeological literature as the "integrated strategy". The application of the method is statistically incorrect since in archaeological studies the hypothesis of independence between sites, which is at the basis of the logistic regression model, is violated.

To overcome such limitations we propose two alternative models. The first one is an autopredictive model in which the probability of ASL is modelled as a function of the observations coming from field surveys in neighbouring zones. This approach accounts for the problem of non-independency of observations, but neglets *a priori* auxiliary information on the archaeological area. The second approach is a more comprehensive one which overcomes the problems of logistic regressions while preserving the role of *a priori* information.