09 | PREDICTIVE MODELLING

Philip Verhagen

Testing Archaeological Predictive Models: A Rough Guide

Abstract: Archaeological predictive modelling is an essential instrument for archaeological heritage management in the Netherlands. It is used to decide where to do archaeological survey in the case of development plans. However, very little attention is paid to testing the predictions made. Model quality is established by means of peer review, rather than by quantitative criteria. In this paper the main issues involved with predictive model testing are discussed. The potential of resampling methods for improved predictive model quality is investigated, and the problems associated with obtaining representative test data sets are highlighted.

Introduction

Archaeological predictive modelling has been embraced for over 15 years as an indispensable tool for archaeological heritage management in the Netherlands. Predictive maps determine where to do archaeological survey when a development plan is threatening to disturb the soil. Despite this general acceptance of the use of predictive modelling for archaeological heritage management purposes, there is a fundamental problem with the current use of predictive models and maps as it is impossible to judge their quality in an objective way. The quality of the models is established by means of peer review, rather than by quantitative methods. Only limited field tests are carried out, and they are not used in a systematic manner to improve the predictive models. Due to this lack of quantitative rigour, both in the model-building as well as in the testing phase, we cannot adequately assess the archaeological and financial risks associated with making a decision on where to do survey. Furthermore, no in-depth studies on predictive model testing have appeared since the papers published in Judge / Sebastian (1988). Consequently, the research project "Strategic research into, and development of best practice for, predictive modelling on behalf of Dutch cultural resource management" (van Leusen / Kamermans 2005) defined as one of its objectives to analyze the potential of quantitative testing methods for predictive modelling, and to describe the consequences of applying these in practice. The current paper summarizes the results of this study, and presents some of its main conclusions. A more detailed account was recently published by Verhagen (2007).

Defining Predictive Model Quality

At least five criteria for predictive model quality can be given:

- Good models should provide an explanatory framework for the observed site density patterns. Just predicting a high, medium or low probability is not enough. We should also know why the prediction is made. In practice, this means that so-called "inductive" predictive models will never be satisfactory (see Wheatley 2003; Whitley 2004).
- Good models should be transparent. The model-building steps should be clearly specified, and the results should be reproducible.
- Good models should give the best possible prediction with the available data set. This means that the models have to be optimized.
- Good models should perform well in future situations. This implies that independent testing is an indispensable part of establishing model quality

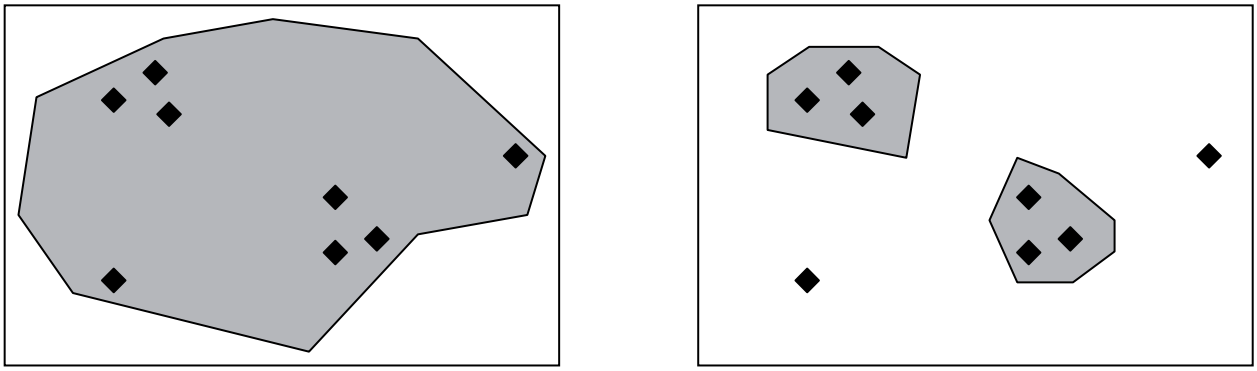

Fig. 1. The differenc between accuracy and precision. The model to the left is 100% accurate: it captures all sites (the black lozenges) in the high probability zone of the model (depicted in grey). The model to the right is less accurate, but more precise.

- Good models should specify the uncertainty of the predictions. This is necessary to establish the risk involved with classifying zones into high, medium or low probability.

As this paper is dealing with the quantitative aspects of model testing, it will not go into detail about the first two criteria involved. The necessity of a satisfactory explanatory framework for a predictive model is generally recognized by archaeologists, even when in theory good predictive models might be produced with "blind" inductive modelling. Similarly, a transparent account on the way in which the model is built is part of the normal scientific process, and should not be a problem in practice.

In most published accounts, predictive model quality is judged by establishing its "performance". This is usually understood to mean a combination of the model's accuracy and precision. Accuracy is equivalent to correct prediction: are most of the sites captured in the high probability zone of the model? Precision refers to the ability of the model to limit the area of high probability as narrowly as possible *(Fig. 1)*.

A predictive model will only be useful for archaeological heritage management purposes when it combines a high accuracy with a high precision. Kvamme's gain[1] is often used to measure model performance, as it combines the two criteria of accuracy and precision in one easily calculated measure. However even when only using gain as a measure of model performance, we are already confronted with the problem of deciding whether the model is good enough. For example, equal gain values can be obtained with different values for accuracy and precision. A 0.5 Kvamme's gain can be reached by including 60% of the sites in 30% of the area, or by including 80% of the sites in 40% of the area. Hence we can define an additional criterion for model quality: does it achieve the goals set by either authorities or developers? Surprisingly enough, these goals have hardly figured in discussions on predictive model quality in the Netherlands. An analysis of the performance of the Indicative Map of Archaeological Values of the Netherlands (Deeben et al. 2002; *Fig. 2*) showed that Kvamme's gain values range from 0.2 to 0.79 for different regions. And when the province of Limburg wanted to know whether it really had to protect 70% of its territory by means of an obligation to do survey, no attention at all was paid to the archaeological risks involved – nor did the financial risks play a major role either. The question therefore is: can we actually establish these risks?

Getting the Best Possible Model

One way of dealing with the risks involved is by optimizing the predictive model. This implies finding the best possible trade-off between accuracy and precision. A class boundary has to be established between the high probability and low probability areas. As low probability implies that no archaeological interventions are necessary, it is important to find the best possible compromise. By shifting class boundaries, accuracy and precision can be changed, but increasing the model's accuracy implies reducing its precision and vice versa. Kvamme (1988) developed the intersection method to find the optimal trade-off between the two, but other methods like gain development graphs have been used as

[1] Specified as $1 - p_a/p_s$, where p_a=the proportion of area (precision) and p_s = the proportion of sites (accuracy) covered by the tested probability zone of a predictive model (Kvamme 1988)

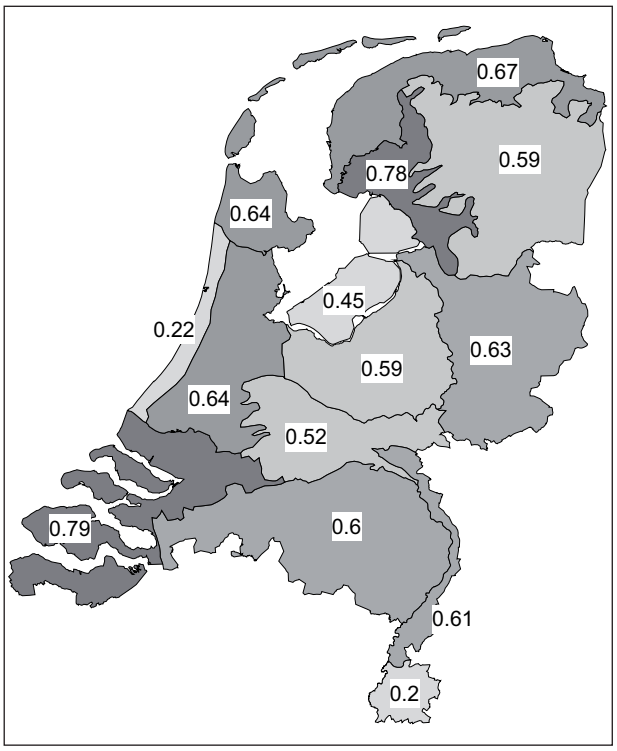

Fig. 2. Kvamme´s gain values for the Indicative Map of Archaeological Values.

well (Deeben et al. 1997; Verhagen / Berger 2001). Optimization is independent of the modelling procedure used, as it is only a way of deciding where to place class boundaries. By imposing thresholds on minimum accuracy and precision we might only be able to control the risks involved to some extent, but it can also provide us with a baseline to compare between different predictive maps.

Establishing Model Error with Resampling

However, when we have established accuracy and precision, we're only half way there. We need to have some idea of the uncertainty of the prediction involved as well. While the optimization of a predictive model might lead to the outcome of say 70% of the known archaeological sites being found within the high probability zone, this does not necessarily mean that this will be true for future cases. In all predictions and classifications, there is an error involved, and the larger this error is, the less useful our model will be. An early concern of predictive modellers therefore was the establishment of measures of the error of their predictions. This is not an easy thing to do, as it all depends on the availability of archaeological test data that constitute a representative sample of the study area. This is the primary reason why deductive modelling is to be preferred to inductive modelling. Even though deductive models are based on subjective weighting, we can always keep the archaeological data apart, and use it for testing purposes. With inductive models we don't have this option, so authors like Rose and Altschul (1988) and Kvamme (1988; 1990) developed several methods for simple validation (or internal testing) of inductive models. These methods were primarily intended to come up with a more realistic estimate of the classification error, while still using the model design data set. Simple validation methods however have not met with general approval in predictive modelling literature. Ebert (2000) for example stated that they are "a grossly inefficient way to determine if there is inhomogeneity in one's data", and Gibbon (2002) noted that all testing methods that use the data from which the model was derived have severe drawbacks (see also Rose / Altschul 1988).

The first option to be used for simple validation is split sampling. It withholds data from the available sample (usually 50%) to see whether the model is capable at predicting the data that is left out from model building. However, split sampling is not very useful for validation purposes, for two reasons. On the one hand, the split sample is not a truly independent sample, as it derives from the data set originally collected for model building. Only if we are sure that these original data was collected according to the principles of probabilistic sampling, can we consider the split sample to be an independent test data set. On the other hand, we should always expect the model to show poorer performance with the split sample than with the design data set, as an inductive model will be optimized to this design set (Hand 1997). And since the stability of models based on small data sets will always be less than the stability of models based on large data sets, it is strongly recommended that the full data set is used for model building, especially since we now have much stronger internal testing methods available in the form of resampling.

Resampling techniques re-use parts of the complete data set in order to obtain a better estimate of the model's error. The simplest resampling method available is cross-validation[2]. It refers to dividing the sample into a number of randomly chosen, roughly equal-sized subsets. Each subset is withheld from the analysis in turn, and a model is developed with the remainder of the data. The withheld subset is then classified using this model, and this is repeated

until all subsets have been used. The total error rate is then determined by averaging the error rates of the subset classifications across the models. Cross-validation used in this way produces a less biased estimate of the true error rate (Hand 1997).

Cross-validation can be taken to extremes by withholding one observation at a time. This is also known as the "leave-one-out" (LOO) approach, and is comparable to what is generally known as jackknife sampling[3]. This method has already been used by others (Rose / Altschul 1988; Kvamme 1988; 1990) to improve their predictive models. The final option to calculate error rates is by means of bootstrap sampling. Unlike jackknife sampling and cross-validation, bootstrap sampling does not divide the data set in a predefined number of subsets, but instead picks a random sample with replacement of size equal to the complete data set (so individual observations may be found in the "subset" more than once; Hand 1997). A model is developed with each subset, and the error rate is determined at each analysis by using the complete data set (which of course contains no double observations). Current statistical opinion favours bootstrapping over jackknife sampling (see Efron / Tibshirani 1993).

The doubts expressed on the utility of simple validation methods for predictive modelling have more to do with a distrust of the data sets used for model building, than with the applicability of the validation methods themselves. Statisticians are quite clear that the application of resampling methods is good practice when it comes to estimating classification error, so resampling (and especially bootstrapping) can be a valuable technique to obtain error estimates for a predictive model, and it is equally applicable to deductive models. Resampling is currently also positioned as an alternative to classical statistical inference by some authors (e.g. Simon 1997). Lunneborg (2000) mentions a number of limitations of classical statistical (parametric) inference. Small sample size, small population size and the assumption of random sampling limit the application of standard statistical inference techniques. Resampling will in those cases generally offer better estimates of the population characteristics than classical inference methods, which rely heavily on the assumption of idealized statistical distributions. This means it can also become of interest in the development of site density estimates and associated confidence intervals as well, provided we have control over the surveyed areas.

Obtaining Independent Test Data

As noted, the best way to test a predictive model is by using an independent, representative data set. The testing method itself is irrelevant to this principle. Whether the independent data set is obtained by keeping data apart, or by collecting new data, it is imperative that the control data is a representative sample of the archaeological phenomena that we are trying to predict. In other words, we have to make sure that a data set of sufficient size is obtained through the principles of probabilistic sampling. This means that the following conditions should be met for independent data collection:

- the sample size should be large enough to make the desired inferences with the desired precision;
- the sampled areas should be representative of the study region;
- survey methods should be chosen such that bias in site recording is avoided.

An important obstacle to this is the difficulty of collecting data from many small survey projects, such as those usually found in archaeological heritage management. The number of sites identified in an individual small survey project will be very limited, so data from various surveys will have to be combined in order to obtain a sufficiently large test set. This not only implies collecting data from different sources, but also of varying quality, which will make it difficult to compare the data sets. There is also a strong possibility that the survey data will not be representative. Low probability areas for example tend to be neglected because the model indicates that there will be no sites (see e.g. Griffin / Churchill 2000 for an example from practice; Wheatley 2003 for a critique of this approach; and Verhagen 2005 for some cases of institutionalized bad habits).

Nevertheless, it seems a waste of data not to use "compliance" survey data for independent testing, especially since it is a data source that has been growing rapidly and will continue to do so. How-

[2] Also known as rotation (Hand 1997); split sampling is sometimes also referred to as cross-validation, but this is not a correct use of the terminology. Baxter (2003) remarks that the term hold-out method is to be preferred for split sampling

[3] However, jackknife error estimation deals somewhat differently with establishing the error rate (see Hand 1997)

ever, there are some statistical and practical difficulties involved in establishing the actual amount of data needed for predictive model testing purposes. The standard procedures to calculate appropriate sample sizes can be found in any statistical handbook (e.g. Shennan 1997; Orton 2000), but these are based on the assumption that samples consist of two classes, like site presence-absence counts per area unit. While the "classical" American logistic regression models are based on site/non-site observations in survey quadrats, in many other studies we are usually dealing with point observations of sites: samples with only one class. Furthermore, most of the time we don't know the proportion of the area sampled, which makes it impossible to specify statistical estimates and corresponding confidence limits of site density. Added to this, we cannot predict the size of the area that should be sampled in order to obtain the required sample size, as long as we don't know the real site density in the survey area. This clearly points to the importance of making models that specify statistical estimates of site density and confidence limits based on probabilistic sampling. It appears we could use resampling techniques to make these calculations.

There are other sampling issues that must be taken into account as well, and especially the influence of survey bias. Unfortunately, methods and procedures for controlling and correcting survey bias have not featured prominently in or outside predictive modelling literature. The main sources of bias identified are:

- the presence of vegetation, which obscures surface sites,
- sediment accumulation, which obscures subsurface sites
- sampling layout, which determines the number and size of the sites that may be found,
- sub-surface sampling unit size, which determines if sites may be detected,
- survey crew experience, which determines if sites are actually recorded.

Orton (2000) mentions imperfect detectability as the main source of non-sampling error in archaeological survey. In theory, correcting site density estimates for imperfect detectability is relatively easy. The task of bias correction becomes a question of estimating the detection probability of a particular survey. Obviously, this would be easiest if survey results were based on the same methods. With this not being the case, a straightforward procedure for bias reduction is to sub-divide the surveys into categories of detectability that can be considered statistical strata. For example, one stratum may consist of field surveys carried out on fallow land with a line spacing of 10 m, a second stratum of core sampling surveys using a 40 x 50 m triangular coring grid and 7 cm augers up to 2 m depth. For each of these categories, site density estimates and variances can be calculated, and must be corrected for imperfect detectability. The calculation of the total mean site density and variance in the study area can then be done with the standard equations for stratified sampling. Even though the procedure is straightforward, this does not mean that the estimation of detection probability is easy. For example, some sites may be characterized by low numbers of artefacts but a large number of features. These will be extremely hard to find by means of core sampling but they do stand a chance of being found by means of field survey if the features are (partly) within the plough zone; and they will certainly be found when digging trial trenches. A quantitative comparison of the success or failure of survey methods is therefore never easy, and very much depends on the information that we have on the prospection characteristics of the sites involved.

In practice, obtaining these may be an insurmountable task. Tol et al. (2004), who set out to evaluate the process of archaeological core sampling survey in the Netherlands and compare it to archaeological excavation, were forced to conclude that this was impossible within the constraints of their budget. This was not just a question of incompatibility of data sources, but also of a lack of clearly defined objectives for prospection projects. Consequently, the survey methods could not be evaluated for their effectiveness. However, in the context of predictive model testing, an alternative could be found by settling for comparable surveys that are adequately described, analysing if there are any systematic biases that need to be taken into account, and using this data as the primary source for retrospective testing. This obviously implies that the factors that influence detection probability should be adequately registered for each survey project. This is far from common practice.

Registration of the fieldwork projects in the Dutch national archaeological database ARCHIS for example turns out to be erratic in the definition of the studied area and the research methods applied. It is impossible to extract the information needed for an analysis of detection probabilities from the

database. Furthermore, a major problem with the delimitation of the surveyed areas is apparent. The municipality of Het Bildt (province of Friesland) contains 26 database entries, covering the entire municipality, and the neighbouring municipality of Ferwerderadeel has another 34. These 60 projects together take up 62.5% of the total registered area of completed research projects. However, most of the 60 entries refer to small core sampling projects, carried out within the municipalities' boundaries, but without any indication of their precise location. Clearly, the fieldwork data in ARCHIS in its current form is not even suitable for rigorously quantifying the bias of archaeological fieldwork to zones of high or low probability. We are therefore forced to return to the original research project documentation to find out which areas have actually been surveyed, and which methods have been applied.

Conclusions and Recommendations

Testing of predictive models is an issue that is far from trivial. We are dealing with a problem that is closely related to the very principles of statistical inference and sampling. Without representative samples, our predictions will always be flawed, no matter whether we are building inductive or deductive models. Methods and procedures for dealing with biased data are still underdeveloped, even though statistical rigour is now somewhat relaxed by the development of resampling techniques. A fundamental hindrance to predictive model testing is found in the fact that standard survey procedures do not incorporate the specification of the factors influencing site detection. Furthermore, the current state of predictive modelling, at least in the Netherlands, does not allow us to clearly define the amount of data that has to be collected in order to achieve the desired model quality. Since the models are not cast into the form of statistical estimates of site density, it is impossible to specify the models' current statistical precision, their desired precision, and the resulting necessary sample size to arrive at this desired precision. At the current state of affairs, the maximum result that can be attained is an estimate of the model's accuracy and precision, based on unevenly documented compliance survey data sets.

The recommendations resulting from this study are therefore straightforward: in order to seriously test predictive models, we should be using statistical estimates and confidence limits instead of pleasantly vague classes of high, medium and low 'probability'. While the currently available survey documentation may yield sufficient representative data (after analysis and correction of survey bias), it is also necessary that future survey campaigns should better take into account the principles of probabilistic sampling. This implies for example that, for testing purposes, low probability areas should be surveyed as well. Furthermore, to reduce the archaeological risk of development plans, clear norms should be defined for the accuracy and precision of predictive models. It is only then that predictive models will become useful tools for quantitative risk assessment.

References

Baxter 2003

M. J. Baxter, Statistics in Archaeology (London 2003).

Deeben et al. 1997

J. Deeben / D. Hallewas / J. Kolen / R. Wiemer, Beyond the Crystal Ball: Predictive Modelling as a Tool in Archaeological Heritage Management and Occupation History. In: W. Willems/ H. Kars / D. Hallewas (eds.), Archaeological Heritage Management in the Netherlands: Fifty Years State Service for Archaeological Investigations (Amersfoort 1997) 76–118.

Deeben / Hallewas / Maarleveld 2002

J. Deeben / D. P. Hallewas / T. Maarleveld, Predictive Modelling in Archaeological Heritage Management of the Netherlands: the Indicative Map of Archaeological Values (2nd Generation). Berichten van de Rijksdienst voor het Oudheidkundig Bodemonderzoek 45, 2002, 9–56.

Ebert 2000

J. Ebert, The State of the Art in "Inductive" Predictive Modeling: Seven Big Mistakes (and Lots of Smaller Ones). In: K. Wescott / R. Brandon (eds.), Practical Applications of GIS For Archaeologists. A Predictive Modeling Kit (London 2000) 129–134.

Efron / Tibshirani 1993

B. Efron / R. J. Tibshirani, An Introduction to the Bootstrap. Monographs on Statistics and Applied Probability 57 (New York 1993).

Gibbon 2002

G. Gibbon, A Predictive Model of Precontact Archaeological Site Location for the State of Minnesota. Appendix A: Archaeological Predictive Modelling: An Overview (Saint Paul 2002). http://www.mnmodel.dot.state.mn.us/chapters/app_a.htm. [31 Dec 2007].

Griffin / Churchill 2000
D. Griffin / T. Churchill, Cultural Resource Survey Investigations in Kittitas County, Washington: Problems Relating to the Use of a County-wide Predictive Model and Site Significance Issues. Northwest Anthropological Research Notes, 34:2, 2000, 137–153.

Hand 1997
D. Hand, Construction and Assessment of Classification Rules (Chichester 1997).

Judge / Sebastian 1988
W. Judge / L. Sebastian, Quantifying the Present and Predicting the Past: Theory, Method, and Application of Archaeological Predictive Modelling (Denver 1988).

Kvamme 1988
K. Kvamme, Development and Testing of Quantitative Models. In: W. Judge / L. Sebastian (eds.), Quantifying the Present and Predicting the Past: Theory, Method, and Application of Archaeological Predictive Modelling (Denver 1988) 325–428.

Kvamme 1990
K. L. Kvamme, The Fundamental Principles and Practice of Predictive Archaeological Modelling. In: A. Voorrips (ed.), Mathematics and Information Science in Archaeology: A Flexible Framework. Studies in Modern Archaeology (Bonn 1990) 257–295.

van Leusen / Kamermans 2005
M. van Leusen / H. Kamermans, Predictive Modelling for Archaeological Heritage Management: A Research Agenda. Nederlandse Archeologische Rapporten 29 (Amersfoort 2005).

Lunneborg 2000
C. E. Lunneborg, Data Analysis by Resampling: Concepts and Applications (Pacific Grove 2000).

Orton 2000
C. Orton, Sampling in Archaeology. Cambridge Manuals in Archaeology (Cambridge 2000).

Rose / Altschul 1988
M. Rose / J. Altschul, An Overview of Statistical Method and Theory for Quantitative Model Building. In: W. Judge / L. Sebastian (eds.), Quantifying the Present and Predicting the Past: Theory, Method, and Application of Archaeological Predictive Modelling (Denver 1988) 173–256.

Shennan 1997
S. Shennan, Quantifying Archaeology (Edinburgh 1997).

Simon 1997
J. L. Simon, Resampling: The New Statistics. http://www.resample.com/content/text/index.shtml [31 Dec 2007].

Tol et al. 2004
A. Tol / P. Verhagen / A. Borsboom / M. Verbruggen, Prospectief boren: Een studie naar de betrouwbaarheid en toepasbaarheid van booronderzoek in de prospectiearcheologie. RAAP-rapport 1000 (Amsterdam 2004).

Verhagen / Berger 2001
P. Verhagen / J. Berger, The Hidden Reserve: Predictive Modelling of Buried Archaeological Sites in the Tricastin-Valdaine region (Middle Rhône Valley, France). In: Z. Stančič / T. Veljanovski (eds.), Computing Archaeology for Understanding the Past - CAA 2000. Computer Applications and Quantitative Methods in Archaeology, Proceedings of the 28th Conference, Ljubljana, April 2000. BAR International Series 931 (Oxford 2001) 219–232.

Verhagen 2005
P. Verhagen, Archaeological Prospection and Archaeological Predictive Modelling. In: M. van Leusen / H. Kamermans (eds.), Predictive Modelling for Archaeological Heritage Management: A Research Agenda. Nederlandse Archeologische Rapporten 29 (Amersfoort 2005) 109–122.

Verhagen 2007
P. Verhagen, Predictive Models put to the Test. In: P. Verhagen (ed.), Case Studies in Archaeological Predictive Modelling, Archaeological Studies Leiden University 14 (Leiden 2007) 115–168.

Wheatley 2003
D. Wheatley, Making Space for an Archaeology of Place. Internet Archaeology 15. http://intarch.ac.uk/journal/issue15/wheatley_index.html [31 Dec 2007].

Whitley 2004
T. G. Whitley, Causality and Cross-purposes in Archaeological Predictive Modeling. In: A. Fischer Asserer / W. Börner / M. Goriany / L. Karlhuber-Vöckl (eds.), Enter the past: the E-way into the Four Dimensions of Cultural Heritage. Computer Applications and Quantitative Methods in Archaeology 2003. BAR International Series 1227 (Oxford 2004) 236–239.

Philip Verhagen

Faculty of Archaeology, ACVU-HBS
Vrije Universiteit
1081 HV Amsterdam, The Netherlands
jwhp.verhagen@let.vu.nl