The Bumpy Road to Incorporating Uncertainty in Predictive Modelling

Philip VERHAGEN[1] – Martijn van LEUSEN[2] – Benjamin DUCKE[3] – Andrew MILLARD[4] – Hans KAMERMANS[5]

[1]ACVU-HBS
[2]Groningen University
[3]Oxford Archaeological Unit
[4]Durham University
[5]Leiden University
[1]jwhp.verhagen@let.vu.nl
[2]p.m.van.leusen@let.rug.nl
[3]benjamin.ducke@oxfordarch.co.uk
[4]a.r.millard@durham.ac.uk
[5]h.kamermans@arch.leidenuniv.nl

Abstract

One of the key problems of predictive modelling is the lack of tools to incorporate and map the uncertainties of the predictions made. Without explicit description of the varying quality of the archaeological and environmental data, statistical methods risk making inaccurate predictions. Hence, lacking adequate descriptions of bias and error, predictions often rely on expert judgement. But can expert judgement be quantified in such a way, that predictions can be made that will respect the experts' views, and at the same time reflect the uncertainties in the experts' opinions as well as in the available data? This paper reports an investigation into whether expert views can be quantified and incorporated in statistical predictions, for which we tested two potentially useful techniques, Bayesian inference and Dempster-Shafer theory.

Keywords

predictive modelling, Bayesian inference, Dempster-Shafer theory, uncertainty

1. Introduction

Anyone working with predictive models knows the slightly uneasy feeling that comes with looking at the brightly or pastel-coloured zones where the probability of encountering archaeological remains is considered to be 'high', 'medium' or 'low'. How can we be so sure that the low probability zones are really not interesting? And where do we draw the line between interesting and not interesting?

Concern over whether predictions can hold in the face of elusive social behaviour, complex geomorphological processes, research biases and data quality has created a painful awareness of the many sources of uncertainty inherent in the models. While we can use the available archaeological data to draw boundaries between high, medium and low probability, this does not tell us whether the predictions are reliable, as long as we can't specify the bias and error in the data set used. And even if we rely on expert judgement for 'correcting' or adjusting predictions, we can expect experts to be uncertain as well, and to disagree among themselves.

Within the research project 'Strategic research into, and development of best practice for, predictive modelling on behalf of Dutch cultural resource management' (van Leusen and Kamermans 2005) we have investigated what methods are best suited for dealing with uncertainty in predictive modelling. For this, we have looked into two relatively new methods for developing predictive models, Bayesian inference and Dempster-Shafer theory. The study region chosen was the Rijssen-Wierden area (*Fig. 1*), where one of the first predictive models in the Netherlands was made (Ankum and Groenewoudt 1990). A more detailed account of the case study will be published in van Leusen *et al.* (2009).

Fig. 1. Location of the Rijssen-Wierden study area in the Netherlands.

2. Bayesian inference

2.1. Introduction

Bayesian inference differs from classical statistics in allowing the explicit incorporation of subjective prior beliefs into statistical analysis (see e.g. Buck *et al.* 1996). This makes it an interesting method for predictive modelling using expert (prior) opinions. A Bayesian statistical analysis produces an assessment of the uncertainty of the calculated probabilities in the form of standard deviations and credibility intervals. It also provides a simple framework for incorporating new data into the model. Bayesian inference, while conceptually straightforward, has only observed widespread application after the advent of powerful computing methods. In archaeology, Bayesian inference is predominantly used in ^{14}C-dating for calibration purposes. In predictive modelling, the number of published applications is limited to two case studies (van Dalen 1999; Verhagen 2006). In addition two other papers (Orton 2000; Nicholson *et al.* 2000) consider survey sampling strategies and the probability that archaeological sites are missed in a survey project, given prior knowledge of site density, such as might be gained from a Bayesian predictive model.

2.2. Application

A Bayesian model was produced for settlement in the study area, showing how conditional probabilities combine with observations to yield posterior probabilities, with an associated measure of uncertainty. To obtain prior probabilities, the experts were asked to rate each of the six 'environmental factors' used in the 1990 model for their relative odds of containing archaeological sites.

Expert 2's odds with regard to the factor soil texture are given in *Table 1*. These relative odds are converted into absolute probabilities ('prior proportions'), summing to 1. Since the expert supplied information on all possible combinations of texture classes, we can make four separate calculations of these prior proportions (the four rows in the right-hand part of the table), which do not necessarily agree. According to the expert, texture class 2 should attract between 17 and 53% of the sites, with a mean of 29% and a coefficient of variance of 56%. In other words, this expert is rather uncertain about some of the odds. The calculated mean provides our best estimate of his true position.

Identical calculations were made for each of the six factors, and for each expert separately. This information was combined to arrive at an assessment of the mean expert opinion and its variance, and the consequent 'data equivalent'. This expresses our reliance on the experts' opinions in terms of the number of actual site observations that would be needed to provide the same amount of information about site distributions.

Table 2 contains the three experts' prior proportions for the factor soil texture, with the corresponding means and standard deviations. The experts are in general agreement about the proportion of sites to be found in each of the soil texture classes, if these were equally represented in the study

	0	**1**	**2**	**3**
0	1	0,5	0,25	0,1
1	2	1	0,25	0,5
2	4	4	1	0,33
3	10	2	3	1

	0	**1**	**2**	**3**	**sum**
	0,059	0,118	0,235	0,588	1
	0,067	0,133	0,533	0,267	1
	0,056	0,056	0,222	0,667	1
	0,052	0,259	0,172	0,517	1
MEAN	**0,058**	**0,141**	**0,291**	**0,510**	1
CV %	**10,9%**	**60,2%**	**56,4%**	**34,0%**	

Table 1. Expert 2's assessment of relative odds for the factor 'soil texture' (left), converted into probabilities (right), with means and variances (bottom). CV = coefficient of variance.

	0	1	2	3	sum
expert 1	0,027	0,051	0,217	0,704	1
expert 2	0,058	0,141	0,291	0,510	1
expert 3	0,150	0,050	0,30	0,500	1
MEAN	**0,079**	**0,081**	**0,269**	**0,571**	**1**
STDEV	**0,064**	**0,052**	**0,046**	**0,115**	

Table 2. Experts' prior proportions for the factor 'soil texture', with means and standard deviations.

area. From this, Dirichlet prior vectors and data equivalents need to be calculated to arrive at the final probability calculations (the Dirichlet-distribution is the conjugate distribution of the multinomial distribution, and the appropriate statistical model for dealing with categorical data like soil classes). Various approaches can be used for this (*Table 3*, methods A–C). In method A each expert is assumed to be worth one observation, and their combined data equivalent is 3. However, a better approach is to find the data equivalent that gives the same standard deviations as the experts' priors, and this is done using methods B1 and B2. Where the experts agree (that is, the standard deviation of their opinions is low) a high data equivalent results; where they disagree a low one results. This is desirable, since we do not value highly the opinion of experts who disagree among themselves, whereas we place more trust in experts who find themselves in agreement. The value α_o is the apparent data equivalent derived from the mean and standard deviation of the experts' opinions for each class. Method B1 uses the mean of these α_o values to arrive at the data equivalent for the factor, whilst method B2 takes a more conservative approach and uses the minimum of the α_o values. So, for the factor soil texture, the experts' priors are calculated to be worth 17 (method B2) or 39 (method B1) observations. We chose to use a data equivalent of 30, as a round figure close to the mean conservative value and typical of the range of values obtained.

	0	1	2	3	data equivalent
method A	0,24	0,24	0,81	1,71	3.0
α_o	16,8	26,1	94,0	17,4	
method B1	3,0	3,1	10,4	22,1	38,6
method B2	1,3	1,4	4,5	9,6	16,8
method C	2,4	2,4	8,1	17,1	30.0

Table 3. Calculation of the experts' data equivalent, using Dirichlet prior vectors. Method A uses a prior "data equivalent" of 3; Method B1 uses the mean α_o from the variance of expert opinions; Method B2 uses the minimum α_o from the variance of expert opinions; and Method C uses a prior "data equivalent" of 30.

This also means that, since we used 80 actual sites for the case study, the experts' opinion represents about a quarter of the weight (30 out of 110) for the final prediction.

Using this calculated 'weight of expert opinion', shown as method C in *Table* 3, the relative probability of finding a site in each of the six 'environmental factors' was calculated (*Fig. 2*). This map summarises the experts' views on the relative density of sites in the landscape. When this is confronted with predictions based on site observations, a number of discrepancies are revealed. We can observe areas where sites are present despite their predicted absence, and areas where sites are absent despite their predicted high density. This is partly as it should be: site discovery is influenced by visibility factors and construction work, so areas with a high site potential that have not been available for research will not have any site observations. Conversely, if a high proportion of sites is found in areas where experts predict they should not be, this must be taken as an indication that either the experts or the base maps, or both, are wrong.

When we now add the site data to the experts' prior predictions, and re-run the model, the *posterior densities* result (*Fig. 3*). The differences between the prior and posterior densities are shown in *Fig. 4*. Incorporating the data has increased the predicted site probability for most of the blue areas in *Fig. 2* (turning them yellow in *Fig. 3*), and generalized somewhat the predictions of zones of relatively high site density (red colours).

The case study demonstrates how quantitative predictive models can be generated on the basis of expert opinion alone, and how a mechanism exists that adapts these models whenever new data become available. Moreover, this approach allows one to manipulate the weight of expert opinion as opposed to the data: in cases where we have poor data but experts we trust, we can assign a high weight to experts' opinions; in cases where we have good data but little expertise we can assign a low weight. Happily, we do not need to be completely subjective in our rating of the quality of our experts: the variation in the expert's opinions itself provides a measurement of uncertainty, which can also be expressed as a map (*Fig. 5*). After again including the observed sites in the model, the uncertainties are shown to change (*Fig. 6*); the difference is depicted in *Fig. 7*.

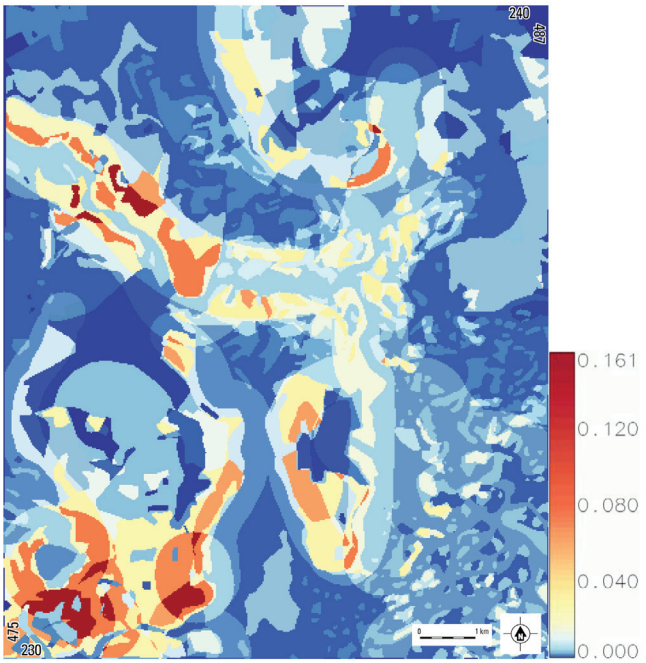

Fig. 2. Relative site density according to experts' prior probabilities. A cell with a value of 0.12 is twice as likely to contain a site as a cell with a value of 0.06.

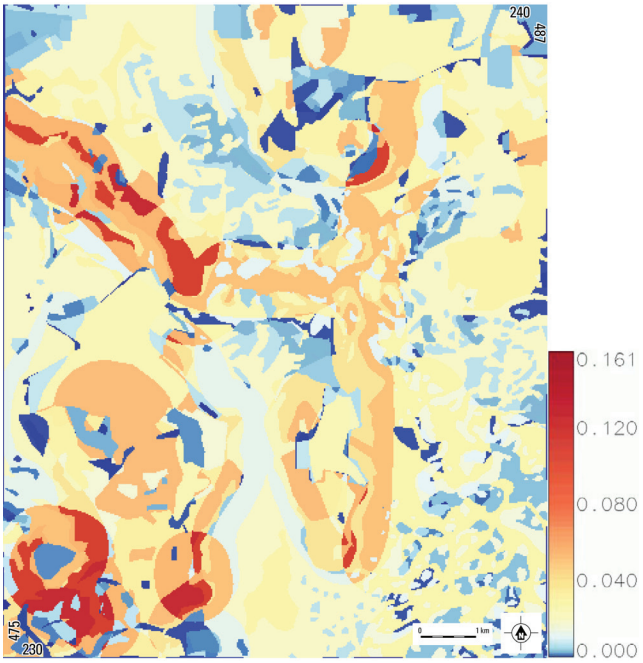

Fig. 3. Posterior site densities after adding site observations to the model.

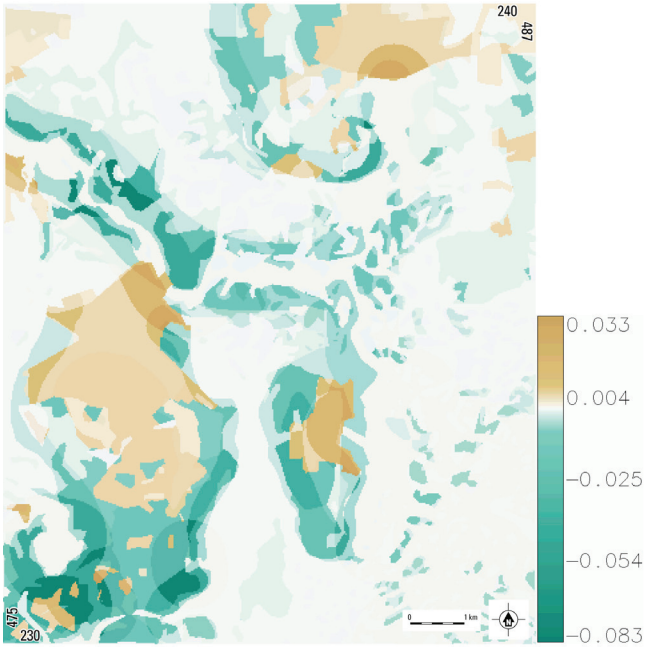

Fig. 4. The difference between figures 2 and 3. Predicted site densities have increased (brown) or decreased (green) when site observations were included into the model.

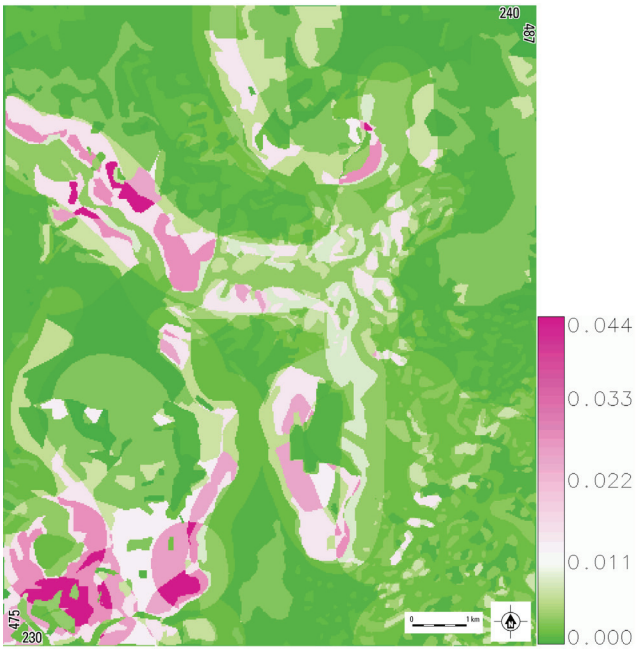

Fig. 5. Uncertainty in the relative densities of sites as modelled by the experts; pink indicates areas of greatest uncertainty. These correspond to the areas of highest density in Fig. 2.

3. Dempster-Shafer Theory

3.1. Introduction

The Dempster-Shafer Theory of evidence (DST) was developed by Dempster (1967) and Shafer (1976), and takes a somewhat different approach to statistical modelling. It uses the concept of *belief*, which is comparable to, but not the same as probability. Belief refers to the fact that we do not have to believe all the available evidence: we can make statements of uncertainty regarding our data. The specification of uncertainty is crucial to the application of DST. Unlike Bayesian inference, DST does not work with

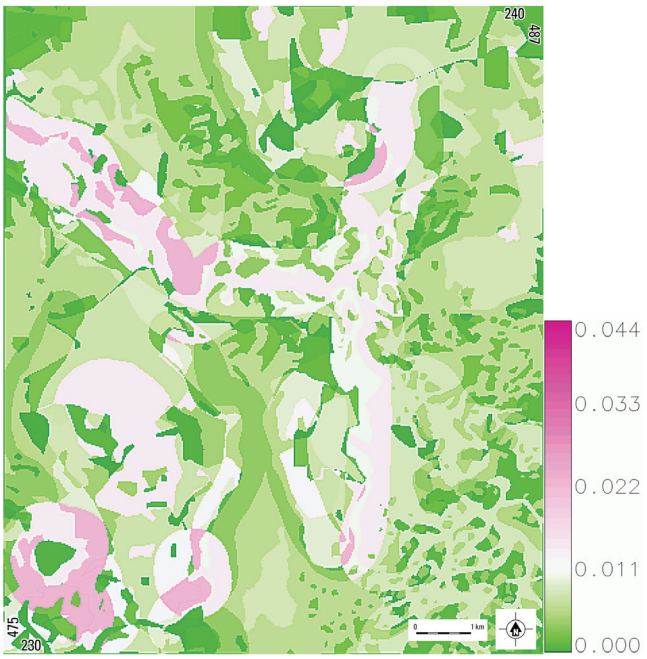

Fig. 6. Uncertainty in the relative densities of sites after the data is included.

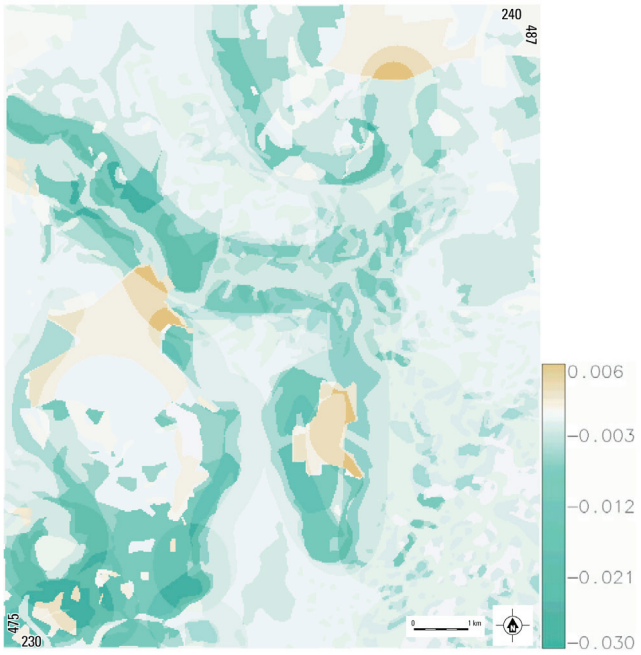

Fig. 7. The difference between figures 5 and 6. Uncertainty has decreased in some areas (green) but increased in others (brown).

an explicit formulation of prior knowledge. Rather, it takes the existing data set and evaluates it for its 'weight of evidence'. The reasons for believing the evidence or not may be of a statistical nature (a lack of significance of the observed patterns, for example), or they may be based on expert judgement (like knowing from experience that forested areas have not been surveyed in the past). DST modelling offers a framework to incorporate these statements of uncertainty. It calculates a measure called *plausibility*, which is the probability that would be obtained if we trust all our evidence. The difference between plausibility and belief is called the *belief interval*, and shows us the uncertainties in the model. Finally, the *weight of conflict* map identifies places where evidences contradict. Different beliefs for different parameters can easily be combined using *Dempsters' rule of combination*.

DST modelling is incorporated in Idrisi and GRASS GIS, and is used for a number of GIS applications outside archaeology. In archaeological predictive modelling, it has been applied in case studies by Ejstrud (2003; 2005). It is better incorporated in GIS and predictive modelling than Bayesian inference. There are clear similarities between DST and (Bayesian) probability theory, as both provide an abstract framework for reasoning using uncertain information. The practical difference is that in a DST model belief values do not have to be proper mathematical probabilities, and much simpler quantifications, such as ratings, may also work (Lalmas 1997).

3.2. Application

In contrast to the Bayesian case study, the DST-modelling did not use the 'old' environmental factors for building the model. The predictive model was built only from data that represents "basic measurements" (e.g. elevation, hydrology) or that has been derived automatically using formalized standard procedures (e.g. slope, aspect, visibility). Some of the original factor maps were produced using weighted overlays and classifications that are highly correlated with the base maps, and this may have introduced an unwanted overweight of certain variables. Because of this, the available "raw" sources of evidence were first analysed for their significance for site distribution, and only the most relevant ones selected for building the model.

In DST-modelling, the first step to be taken then is the establishment of what is called the *basic probability number* (BPN) or *probability mass* of each class in a single map. The BPN expresses the strength of belief in the truth of a hypothesis for a single source of evidence. These BPNs are calculated for two different hypotheses, the {site} and {no site} hypothesis. A calculation of BPNs for all selected sources of evidence supplied ten different "belief maps" for the {site} and {no site} hypotheses respectively. It

is important to keep in mind that the belief outcomes for {site} are not necessarily the inverse of {no site}, as DST-modelling also includes a third hypothesis of uncertainty. If there is insufficient support for either the {site} or {no site} hypotheses, some of the basic probability mass of these hypotheses is transferred to the uncertainty or {site, no site} hypothesis. This was done in either one of the following cases:

- The probability P that the observed difference in proportion between sample (sites) and population (entire region) for an evidence category C could also be produced by chance is > 0. In this case, P is subtracted from the mass for either {site} or {no site} and transferred to the {site, no site} hypothesis for this particular category.
- The chi-square test shows that the overall frequencies of categories in the sample could also have been produced by chance with probability *P*. In this case, *P* is subtracted from the probabilities for either {site} or {no site} and transferred to the {site, no site} hypothesis for *all* categories.
- One or more bias maps are supplied. These specify the degree to which it is believed that observed differences are biased towards the {no site} hypothesis, for example when land use has influenced archaeological discovery rates. For each bias map, the following is done: (a) calculate the percentage of cells *BP* of each category *C* that are covered by a bias value larger than 0; (b) calculate the mean value *BM* of the bias in cells of category *C*. For each category *C*, *BM* * *BP* is subtracted from the mass assigned to {no site} and shifted to the {site, no site} hypothesis.
- One or more attributes of the site vector point map are specified to represent the degree to which these points are biased towards the {site} hypothesis (e.g.: would the presence of a few ceramic sherds be counted as evidence for a site or not?). Calculations are similar to the previous situation. The more biased sites are present on a certain category of an evidence maps, the more mass will be subtracted from the {site} hypothesis and shifted to the {site, no site} hypothesis.

In summary, a high amount of basic probability mass is shifted to the uncertainty hypothesis for category *C*, if (a) many cells of category *C* fall into biased areas and (b) these cells have a high bias on average and/or many sites on category *C* are (strongly) biased. We can then simply combine any number of evidences and their belief mass distributions, including those parts assigned to individual uncertainty hypotheses (*Figs 8* and *9*).

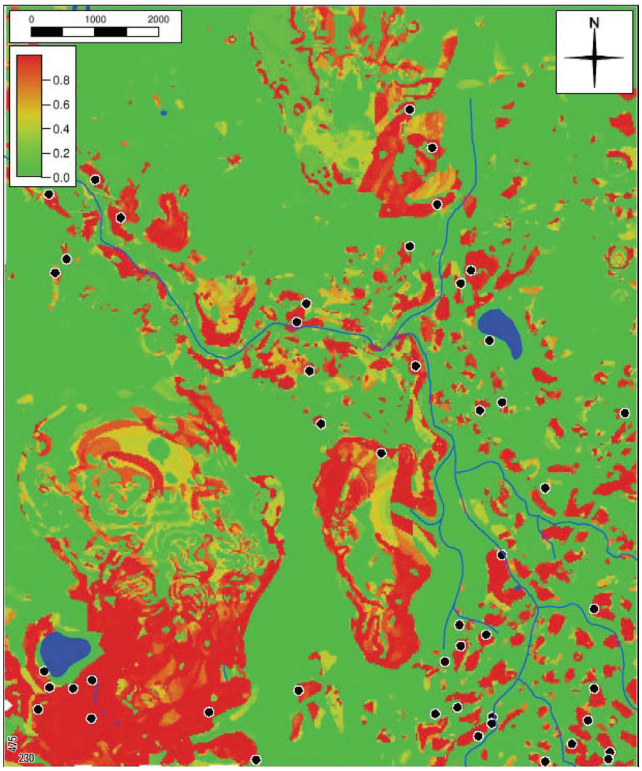

Fig. 8. Map of belief in the {site}-hypothesis for Palaeolithic and Mesolithic sites. Principal lakes and rivers as well as positions of sites used in building the model are indicated as well.

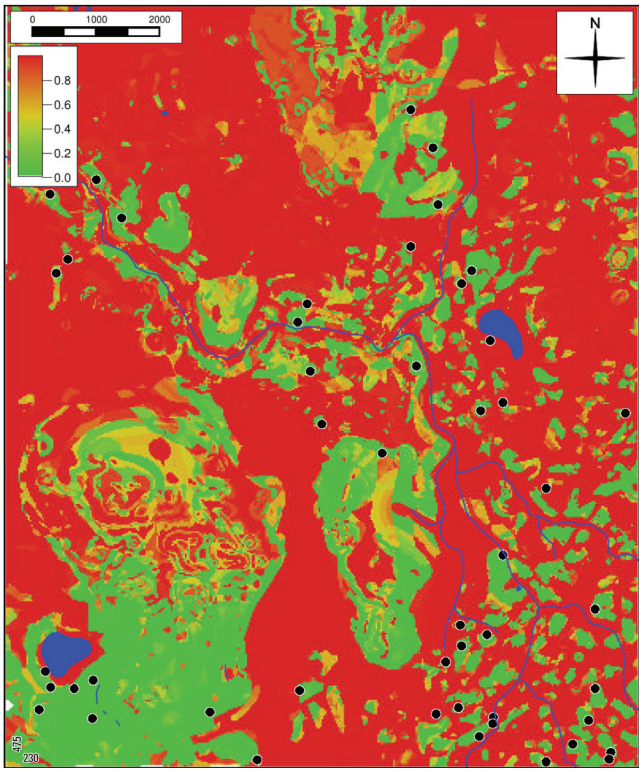

Fig. 9. Map of belief in the {no site}-hypothesis for Palaeolithic and Mesolithic sites. Principal lakes and rivers as well as positions of sites used in building the model are indicated as well.

While the role of experts in setting up the model is restricted, they can play an important role in creating the bias maps. For example, a land use map could be rated by the archaeological experts involved for its contribution to survey bias. Obviously, agricultural land has a much higher probability of revealing archaeological sites during field survey than forest or heather, but a statistical analysis of this effect would be very difficult (see also Verhagen 2007, 146–152). In this case, using expert ratings is an acceptable and much easier solution.

4. Conclusions

The results of the modelling exercises show that Bayesian inference and DST modelling are both capable of including and visualizing uncertainty in predictive modelling. Because the DST modelling in this case study used different environmental factors than the Bayesian modelling, we could not perform a direct comparison between the two. We can however assume that even with a comparable input, the results of the methods will be different, which brings with it the question what will be the best approach. The answer to that question should consider practical issues of versatility, robustness, computational performance and interpretability of model results more than mathematical accuracy, as the latter is adequate in both cases.

Given the preference of DST modelling for using existing data sets instead of formulating prior knowledge, we can assume that Bayesian modelling will be the most appropriate when few data are available. It will then show us where the experts are uncertain, and this could imply targeting those areas for future survey. Bayesian modelling however does not supply a clear mechanism for dealing with (supposedly) unreliable data, while the DST approach implements this by simply stating that these data can only partially be trusted, and hence will only have a limited effect on the modelling outcome.

Getting the required information for Bayesian modelling out of the experts can be somewhat of a struggle, as they are asked to quantify aspects which they are used to thinking about in qualitative mode. It should be stressed that the amount of disagreement displayed by multiple experts provides a relatively objective measure of uncertainty. This also introduces the question of the *relative expertise* of the experts, as we also need a mechanism to rate the accuracy of their opinions.

Predictive models should also be *updatable* with new factor maps, archaeological observations, and expertise. Additional archaeological observations in both approaches are simply used to update the model, but if a weight assigned by an expert changes, or a new type of expertise is added, then the whole model must be recalculated. Additional factor maps require new expertise to be generated, hence also lead to a full recalculation in both approaches. If factor maps are only changed (e.g. a finer resolution soil map becomes available), then the model can be simply updated.

For practical purposes, the results of the models will have to be translated into clear-cut zones. In a simple matrix (*Fig. 10*) the possible 'states' of the model can be shown, with 9 different combinations of predicted site density and uncertainty. For end users of the models, who have to decide on the associated policies, this means that the number of available choices increases from 3 to 9. A reduction to 4 categories might therefore be preferable, only distinguishing between high and low site density and uncertainty. After all, why do we still need the medium class? Usually, this is the zone where we 'park' our uncertainties, so a binary model plus an uncertainty model should do the same trick. The end users then only need to specify how (un)certain they want the prediction to be.

Acknowledgements

We want to thank the archaeological experts involved for their willingness to participate in the case studies: Bert Groenewoudt and Roy van Beek (Rijksdienst voor Archeologie, Cultuurlandschap en Monumenten, Amersfoort) and Huub Scholte Lubberink (RAAP Archeologisch Adviesbureau BV, Brummen).

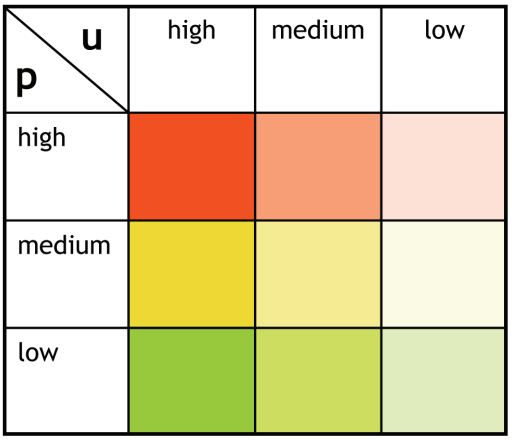

Fig. 10. Simplified scheme for representing predicted site densities (p) and uncertainty (u) in predictive mapping.

References

Ankum, L. Alfred and Bert J. Groenewoudt (1990). *De situering van archeologische vindplaatsen*. (RAAP-rapport 42) Amsterdam: Stichting RAAP.

Buck, Caitlin E., William G. Cavanagh and Clifford D. Litton (1996). *Bayesian Approach to Interpreting Archaeological Data*. Chichester: John Wiley and Sons.

Dalen, Jan van (1999). Probability modelling: a Bayesian and a geometric example. In: Mark Gillings, David Mattingley and Jan van Dalen (eds.) *Geographical Information Systems and Landscape Archaeology*. (The Archaeology of Mediterranean Landscape 3) Oxford: Oxbow Books, 117–124.

Dempster, Arthur P. (1967). Upper and lower probabilities induced by a multivalued mapping. *Ann. Math. Statist.* 38, 325–339.

Ejstrud, Bo (2003). Indicative Models in Landscape Management: Testing the Methods. In: Jürgen Kunow and Johannes Müller (eds.) *Symposium The Archaeology of Landscapes and Geographic Information Systems. Predictive Maps, Settlement Dynamics and Space and Territory in Prehistory. Forschungen zur Archäologie im Land Brandenburg 8. Archäoprognose Brandenburg I.* Wünsdorf: Brandenburgisches Landesamt für Denkmalpflege und Archäologisches Landesmuseum, 119–134.

Ejstrud, Bo (2005). Taphonomic Models: Using Dempster-Shafer theory to assess the quality of archaeological data and indicative models. In: Martijn van Leusen and Hans Kamermans (eds.), *Predictive Modelling for Archaeological Heritage Management: A research agenda*. (Nederlandse Archeologische Rapporten 29) Amersfoort: Rijksdienst voor het Oudheidkundig Bodemonderzoek, 83–194.

Leusen, Martijn van and Hans Kamermans (eds.) (2005). *Predictive Modelling for Archaeological Heritage Management: A research agenda*. (Nederlandse Archeologische Rapporten 29) Amersfoort: Rijksdienst voor het Oudheidkundig Bodemonderzoek.

Leusen, Martijn van, Andrew R. Millard and Benjamin Ducke (2009). Dealing with uncertainty in archaeological prediction. In: Hans Kamermans, Martijn van Leusen and Philip Verhagen (eds.) *Archaeological Prediction and Risk Management. Alternatives to current approaches*. (ASLU 17) Leiden: Leiden University Press, 123–-160.

Lalmas, Mounia (1997). Dempster-Shafer's Theory of Evidence Applied to Structured Documents: Modelling Uncertainty. In: SIGIR '97: *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, July 27–31, 1997*. Philadelphia: ACM, 110–118.

Nicholson, Mike, Jon Barry and Clive Orton (2000). Did the Burglar Steal my Car Keys? Controlling the Risk of Remains Being Missed in Archaeological Surveys. Paper presented at the Institute of Field Archaeologists Conference, Brighton, April 2000. (UCL Eprints) London: University College London (http://eprints.ucl.ac.uk/archive/00002738/01/2738.pdf).

Orton, Clive (2000). A Bayesian approach to a problem of archaeological site evaluation. In: K. Lockyear, T. Sly and V. Mihailescu-Birliba (eds.) *CAA 96. Computer Applications and Quantitative Methods in Archaeology*. (BAR International Series 845) Oxford: Archaeopress, 1–7.

Shafer, Glenn (1976). *A Mathematical Theory of Evidence*. Princeton: Princeton University Press.

Verhagen, Philip (2006). Quantifying the Qualified: the Use of Multicriteria Methods and Bayesian Statistics for the Development of Archaeological Predictive Models. In: Mark Mehrer and Konnie Wescott (eds.) *GIS and Archaeological Site Location Modelling*. Boca Raton: CRC Press, 191–216.

Verhagen, Philip (2007). *Case Studies in Archaeological Predictive Modelling*. (ASLU 14) Leiden: Leiden University Press.