Archaeological Prediction and Risk Management. Alternatives to current practice.

Kamermans, H.; Leusen, M. van; Verhagen, P.

ASLU (Archaeological Studies Leiden University) is a series of the Faculty of Archaeology, Leiden University. The first volume of ASLU appeared in 1998. The series' aim is to publish PhD theses and other research of the faculty. Since 2007 the series has been published as a printing on demand service at Leiden University Press.

The Netherlands are one of the few countries in Europe where predictive models play an important role in cultural heritage management. The models are used to predict archaeological site location in order to guide future developments in the modern landscape. Many scholars however consider the application of predictive models for this purpose highly controversial. Between 2002 and 2006 a team of Dutch researchers conducted strategic research into predictive modelling on behalf of Dutch cultural resource management. One of the goals was to develop best practices for the production and application of these models. This book is the second and final edited volume of publications of this Predictive Modelling project. It brings together technical papers on developing new methods for predictive modelling and applied, interdisciplinary 'action research' focusing on how the models are, or should be, used by stakeholders in cultural heritage management in the Netherlands.

Who should read this book? The main beneficiaries should be those who are involved in making and evaluating policies for the management of archaeological heritage: local and regional government planning officers, and external consultants. But also archaeologists, archaeology students, planners and politicians from the Netherlands as well as from other countries can learn from our efforts. Our experiences can, and already do, guide developments in other European countries and the rest of the world.

Leiden University Press

9 789087 280673

LUP

ASLU

17

ARCHAEOLOGICAL PREDICTION AND RISK MANAGEMENT

alternatives to current practice

H. Kamermans, M. van Leusen, Ph. Verhagen (eds)

ARCHAEOLOGICAL PREDICTION AND RISK MANAGEMENT

H. Kamermans *et al.* (eds)

LEIDEN UNIVERSITY PRESS

Archaeological Prediction and Risk Management

Leiden University Press

Archaeological Studies Leiden University
is published by Leiden University Press, the Netherlands

Series editors: C.C. Bakels and H. Kamermans

Cover Design: Medy Oberendorff
Layout: Hans Kamermans and Medy Oberendorff
Illustrations: Joanne Porck and Medy Oberendorff

ISBN 978 90 8728 067 3
e-ISBN 978 90 4851 063 4
NUR 682

ARCHAEOLOGICAL STUDIES LEIDEN UNIVERSITY 17

Archaeological Prediction and Risk Management

Alternatives to Current Practice

Edited by

Hans Kamermans
Martijn van Leusen
Philip Verhagen

Contents

Preface

This is the second and final edited volume of publications of a project called 'Strategic research into, and development of best practice for, predictive modelling on behalf of Dutch cultural resource management'. The stated goals of the project, *to conduct strategic research into predictive modelling on behalf of Dutch cultural resource management*, *and to develop best practice for it*, were first approached broadly in order to set the international research agenda, and then by targeting specific topics and questions of practical interest to the main stakeholders in the management of the Dutch archaeological heritage.

This volume accordingly combines pure research into methods for predictive modelling of the distribution of archaeological remains in the Dutch soil, with an effort at interdisciplinary 'action research' into the use of such models by stakeholders in Dutch cultural heritage management. This latter approach, which lies at the heart of the BBO programme, brings with it a particular set of problems to do with learning to work with people who have very different goals, approaches, and languages from those in the safe academic environment. We would like to acknowledge here their often enthusiastic interest and participation, the insight they have given us both in their interests and in the limitations within which they must work - as well as our own. In particular, we want to thank the participants in the 'uncertainty meeting' of January 2005 and the 2nd project workshop of March 2006. We are grateful to RAAP Archeologisch Adviesbureau (directed by Marten Verbruggen) for hosting the 2005 meeting, to Leiden University for hosting the 2006 meeting and to NWO Geesteswetenschappen for providing the funding which made it all possible (grant no's 014-18-800 and 240-60-007). Last but not least we like to thank Kelly Fennema, Medy Oberendorff and Joanne Porck for their help with the production of this volume.

Who should read this book? The main beneficiaries, we believe, will be those who are involved in making and evaluating policies for the management of our archaeological heritage: local and regional government planning officers, and external consultants. We hope that the State Service for Archaeology, Cultural Landscapes and Built Heritage (RACM)[1], which despite recent changes in the Dutch Monument and Historic Buildings Act is still very influential in promoting the use of predictive models in planning procedures, will respond to the issues raised in this volume. For easy reading, this volume is split into two parts, separating the material regarding the handling of risks associated with the presence of unknown archaeological remains in the realm of spatial planning and heritage management (Part I), from the more technical work on improvements to predictive modelling methodology (Part II).

By 'risk' we mean two types of risk. First the financial risk to the developer: the economic risk. Second the risk of having archaeology destroyed unnoticed: the scientific risk. Apparently the economic risk is not a big problem. Most of the time the costs of archaeological research are small compared with the overall development costs. Most developers do not worry about these costs. The second type of risk is in the perception of archaeologists very important. Missing archaeological information during the course of a project, either by missing the sites or by destruction of find spots, is one of the main reasons why predictive modelling has such a bad name. After all, the most important use of predictive modelling in Dutch cultural heritage management is as an instrument for selection.

The Editors: Hans Kamermans, Martijn van Leusen and Philip Verhagen

---

[1] Since 2006 the Dutch National Service for Archaeological Heritage (ROB) merged into the RACM, de Rijksdienst voor Archeologie, Cultuurlandschap en Monumenten, the National Service for Archaeology, Cultural Landscape and Built Heritage. Throughout this book we will use the name National Service for Archaeology, Cultural Landscape and Built Heritage and the abbreviation RACM.

1. Archaeological prediction and risk management

Hans Kamermans[2], Martijn van Leusen[3] and Philip Verhagen[4]

1.1 Introduction

Since the adoption of the European Convention on the Protection of the Archaeological Heritage on Malta in 1992 (also known as 'the Malta Convention', 'the Convention of Valletta' or 'the Valletta treaty'; Council of Europe 1992), archaeology in the Netherlands has not been the same. From then on, as a direct result of the Convention's aim 'to protect the archaeological heritage as a source of the European collective memory and as an instrument for historical and scientific study', archaeology has played an important role in spatial planning. All over Europe, archaeology is under threat from development plans, and the Convention was drafted to remedy this. Article 5 reads: 'Each Party undertakes: to seek to reconcile and combine the respective requirements of archaeology and development plans by ensuring that archaeologists participate in planning policies designed to ensure well-balanced strategies for the protection, conservation and enhancement of sites of archaeological interest'. Whereas before, archaeology was about the past, it is now about the future.

And all of a sudden there was money for archaeological research. The Convention specifies that financing should be done by 'taking suitable measures to ensure that provision is made in major public or private development schemes for covering, from public sector or private sector resources, as appropriate, the total costs of any necessary related archaeological operations'. Archaeology became 'developer funded'. These changes in the position and funding of archaeologists brought about a profound shift in the archaeological profession. Archaeology was, and is, no longer the playground of a relatively small group of people who want to know about the past: it has become socially relevant, part of a world that is concerned about its future.

1.2 The Dutch situation

Across Europe, countries implemented the Malta Convention in their own specific manner (Willems 2007). Whilst all embrace the 'developer funded' principle, this is about the only thing they have in common. Spatial planning in the Netherlands has a long tradition, starting with the building of the first dikes circa 1000 AD. Being densely populated and under continual threat of flooding from the sea and the rivers, cooperative planning was needed to remain safe and avoid conflicts about the right to use the limited space available. Various governmental bodies were therefore formed very early on to produce spatial plans and to control their implementation. The management of our archaeological heritage has now become an integral part of this system.

Before 'Malta' the Dutch provinces together with the National Service for Archaeology, Cultural Landscape and Built Heritage (RACM) were responsible for the archaeology in the Dutch soil. Finds had to be reported and only government related agencies were allowed to do archaeological research. These agencies were the RACM, the National Museum of Antiquities, the Universities and some municipalities.

After 'Malta' for some years an interim situation existed. During this period developers who were about to disturb the soil were confronted with their responsibilities, archaeological companies were given access to an archaeological market, and the role of the provinces and the RACM became even more important. A whole set of techniques, rules and regulations were developed during this period.

Since the revised Monuments and Historic Building Act was passed (2007) 'Malta' is fully implemented in the Dutch system. Developers have to pay for the archaeological research and the main responsibility for archaeology is now in the hands of municipalities.

Since the 1990s archaeological predictive modelling has been used as a tool in an early stage of the archaeological heritage management cycle in the Netherlands. Predictive Modelling is a technique to predict,

[2] Faculty of Archaeology, Leiden University, the Netherlands.

[3] GIA, Groningen University, the Netherlands.

[4] ACVU-HBS, Amsterdam, the Netherlands.

at a minimum, the location of archaeological sites or materials in a region, based either on the observed pattern in a sample or on assumptions about human behaviour (Kohler and Parker 1986: 400). There are two reasons to apply predictive modelling in archaeology:

- to gain insight into former human behaviour in the landscape; an academic research application;
- to predict archaeological site location to guide future developments in the modern landscape; an archaeological heritage management application.

Predictive modelling plays an important role in desk-based archaeological assessment studies, indicating where as yet unknown archaeological remains in the soil might affect spatial planning. In the Dutch 'post-Malta' archaeological heritage management practice three parties or 'stakeholders' interact in a free market: the developer, the archaeological contractor, and the authorities. The developer needs archaeology to be dealt with; the archaeological contractor wants to do the archaeological research as efficiently as possible, and the authorities determine the desired quality of research, and perform quality control.

Predictive maps, indicating where there is archaeology in the soil or where it can be expected, now play a major role in spatial planning. By avoiding areas with a high 'risk' of archaeology, developers can reduce the costs involved for archaeological research. This is because predictive maps will always carry policy advice. In some areas, developers will not be obliged to do archaeological research. In others, they might avoid this by taking mitigating measures, like not disturbing the soil during building. And in other cases, archaeological research will be inevitable, in some cases leading to very expensive excavations. So one could say that archaeological predictive modelling is a condition without which archaeology could not be part of the spatial planning process. In the Netherlands, predictive maps are produced on two scales: national and regional. The National Service for Archaeology, Cultural Landscape and Built Heritage (RACM) has produced several versions of the Indicative Map of Archaeological Values of the Netherlands (IKAW). Commercial companies produce predictive maps on a regional or local scale – the latter typically covering a single large municipality or several smaller ones.

The fact that predictive maps play a role in this process and the fact that one of the consequences of this process is selection (some areas will be studied, others will not), the quality of the maps is very important since they are involved in what should be called a form of risk management.

1.3 The project

However, academic experts both in Europe and in North America have contested the use of predictive modelling in archaeological heritage management because of its theoretical and methodological shortcomings (cf. Ebert 2000; Wheatley 2003; Woodman and Woodward 2002; an overview can be found in Van Leusen and Kamermans 2005). Throughout the 1990s, the situation was characterised by separate 'academic' and 'management' discourses. In 2001, the newly established national NWO-funded research programme 'Protecting and Developing the Dutch Archaeological-Historical Landscape' (BBO, Bloemers 2001) sponsored a group of Dutch researchers to begin a thorough study of archaeological prediction, and to establish a meaningful link between scientific knowledge, archaeological-historical heritage management and applied planning policy in the Netherlands The official name for the project was 'Strategic research into, and development of best practice for, predictive modelling on behalf of Dutch cultural resource management' (Kamermans *et al.* 2005). This project ran from 2002 to 2006 at the archaeological departments of the universities of Leiden and Groningen. Martijn van Leusen from the University of Groningen was appointed as a post-doc to be the main researcher. Philip Verhagen has a background in commercial archaeology and was paid by the project for several months to finish a chapter for his PhD thesis. Initially three researchers from the RACM, Jos Deeben, Daan Hallewas and Paul Zoetbrood were also involved. Hans Kamermans from Leiden University directed the project.

The first product of the project was the *Baseline Report on Predictive Modelling for Archaeological Heritage Management in the Netherlands* (Van Leusen *et al.* 2002), which summarizes and analyses current

national and international approaches to predictive modelling on the basis of a review of the available literature.
Six areas of concern were identified:

- quality and quantity of archaeological input data;
- relevance of the environmental input data;
- need to incorporate social and cultural input data;
- lack of temporal and/or spatial resolution;
- use of spatial statistics; and
- testing of predictive models.

Extensive comments were provided on the Baseline Report by a range of international experts invited to a meeting held at the offices of the RACM on 22 and 23 May 2003. The full scientific papers submitted by these experts were subsequently published together with the Baseline Report in an edited volume titled *Predictive Modelling for Archaeological Heritage Management: A research agenda* (Van Leusen and Kamermans 2005).

The subsequent research of the group focused mainly on aspects of the first and the two last topics. Martijn van Leusen investigated the role of expert judgement on the quality of archaeological input data (this volume chapter 9), Philip Verhagen examined the use of statistics and testing in predictive modelling (Verhagen 2007, this volume chapter 8). For the third topic, the incorporation of social and cultural data into predictive models, the project tried in vain to find additional funding (Verhagen *et al.* in press). The research of Hans Peeters, although not part of the project, tackled the two remaining topics (Peeters 2007).

A complete list of publications by the research team can be found in the appendix of this chapter.

1.4 This volume

This book however starts with the experiences of the users. On the 1st and 2nd March 2006 a symposium on '*Archaeological Prediction and Risk Management*' was held in the Kamerlingh Onnes Building of Leiden University. Five experts, all professional users of predictive maps, were invited to present their views on the role of predictive models in Dutch archaeological heritage management. This meeting was intended as a form of 'action research', where problems are being solved in a collaborative context with researchers and other stakeholders. The written contributions of the participants, brief outlines of which are given below, form the first part of this volume.

Marten Verbruggen, director of RAAP, writes about the role of predictive models in prospective research in Dutch archaeological heritage management. After a first success of the small-scale and broad inductive predictive maps, the demand is increasing sharply for large-scale, detailed deductive maps. Nowadays municipalities are the main commissioners of these maps, as they have been given far-reaching powers in the field of the heritage management by the recent amendment to the Monuments and Historic Buildings Act. In the course of the more then 15 years that these maps have been around, it has however become clear that the aim with which these maps were made and the role they subsequently played in the decision-making process in the archaeological heritage management have changed radically.

In *The high price or the first prize for the archaeological predictive model* Martin Meffert, archaeologist for the province of Noord-Brabant, stresses the point that archaeological predictive models form the basis of practically all archaeological research in the Netherlands. He considers the production of predictive maps by archaeological companies undesirable because they could potentially profit from defining large areas with a high indicative value. The manufacturing of predictive maps should lie in the hands of a government institution that is able to operate independently of the market. And since most large university archaeological institutes have their own archaeological company that is dependent on commissions from the market, the universities have thus become structurally dependent on this non-government funding. That is the reason why, in Mefferts view, Dutch

universities can no longer take on this independent role either, and the only suitable candidate left would be the RACM. We would qualify Meffert's conclusion on both counts: firstly, methodological research at universities certainly is not market-dependent, and secondly, the RACM as an institution is not really independent from the archaeological market either - it too profits from a broad definition of 'areas of high indicative value'.

In *Archaeology as a risk in spatial planning: manoeuvring between objectivity and subjectivity*, René Isarin, Philip Verhagen and Boudewijn Goudswaard, all working for archaeological companies, highlight some of the risks in present-day Dutch archaeological heritage management. They consider the risk from the viewpoint of the civil contractor or initiator of a specific spatial development, and not as the risk for the archaeological remains in that specific area. Focusing on the risks related to the first phase of archaeological research (assessment studies), they discuss whether solutions for risk management may be found in the use of predictive modelling and in the development of reliable core sampling survey strategies. In the end the goal must be to reduce risk and uncertainty for the initiator to an acceptable level.

Whatever the shortcomings of current predictive models, they have become an accepted instrument in the planning process. Since no formal publications exist on the subject, the best way to assess the stakeholders' (managers and authorities) views is to look at contested situations, where an attempt by one party to limit the rights of another to do damage to archaeological remains is contested in court. This usually revolves around planning permissions; a case brought before the Dutch Council of State by the municipality of Coevorden is used to illustrate the central role played by differences in interpretation of the IKAW.

The second part of this volume presents work by the project team on the two themes selected for in-depth study after the midterm review. The first two chapters discuss the testing of predictive models. The short *rough guide* (chapter 7) was written especially for the non-technical reader while a full discussion of the theme of testing predictive models can be found in chapter 8. The second set of chapters presents new approaches to predictive modelling, in which the relation between expert knowledge and archaeological input data is redefined.

In chapter 7 and 8, Philip Verhagen explains that archaeological predictive modelling is an essential instrument for archaeological heritage management in the Netherlands. It is used to decide where to conduct 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 Verhagen's first chapter, *Testing archaeological predictive models: a rough guide*, the main issues involved with predictive model testing are discussed. The potential of resampling methods for improving predictive model quality is investigated, and the problems associated with obtaining representative test data sets are highlighted. The commissioned chapter *Predictive models put to the test*, published previously as a part of Verhagen's (2007) PhD thesis, investigates in more detail how one could test predictive models in a quantitative manner, using either old or new archaeological data. Both chapters give some serious warnings concerning the current use of predictive models in the Netherlands: without quantitative quality norms, the models will remain uncontrollable.

The project team considered *reasoning with uncertainty* a major research theme for the second phase of the project. A pilot study on this theme was organised from 17th to 21st January 2005 in Amsterdam, in the offices of RAAP Archeologisch Adviesbureau. Two experts in spatial statistics from Germany (Benjamin Ducke) and the UK (Andrew Millard) together with experts in Dutch archaeology (Bert Groenewoudt, Roy van Beek and Huub Scholte Lubberink), tried to improve the predictive models for a study area in the eastern part of the Netherlands with the application of Bayesian inference statistics and Dempster-Shafer belief and plausibility models. Martijn van Leusen, Benjamin Ducke and Andrew R. Millard publish the results of this study in chapter 9: *Dealing with uncertainty in archaeological prediction*. They created a worked example based on a widely published Dutch case study (Rijssen-Wierden) (Ankum and Groenewoudt 1990; Brandt *et al.* 1992), in such a way that the benefits of these new approaches could be made clear to non-technical readers as well as to those working in heritage management positions. The chapter aims to demonstrate that these approaches can result in useful models for CRM decision support (hopefully more useful than the existing traditional models), and

that concepts of expertise and uncertainty can be given a place in formal models without compromising on robustness and replicability, and hence can provide the link to concepts of (financial) *risk* employed by the people who use these models in the real world.

All these contributions show the application of predictive modelling in archaeology is a fascinating one but not without its problems. It is clear that there is a future for the use of this technique in archaeological heritage management.

References

Ankum, L.A. and Groenewoudt, B.J. 1990. *De situering van archeologische vindplaatsen*. RAAP-rapport 42. Amsterdam: Stichting RAAP

Brandt, R., B.J. Groenewoudt and K.L. Kvamme 1992. An experiment in archaeological site location: modeling in the Netherlands using GIS techniques. *World Archaeology* 2: 268-282

Bloemers, J.H.F. 2001. Het NWO-Stimuleringsprogramma 'Bodemarchief in Behoud en Ontwikkeling' en de conceptuele studies. In J.H.F. Bloemers, R. During, J.H.N. Elerie, H.A. Groenendijk, M.Hidding, J. Kolen, Th. Spek, and M.-H. Wijnen (eds), *Bodemarchief in Behoud en Ontwikkeling. De Conceptuele Grondslagen*, 1-6. Den Haag: NWO

Council of Europe, 1992. *European Convention on the Protection of the Archaeological Heritage (Revised)*. European Treaty Series 143

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

Kamermans, H., J. Deeben, D. Hallewas, P. Zoetbrood, M. van Leusen and Ph. Verhagen 2005. Project proposal. In M. van Leusen and H. Kamermans (eds), *Predictive Modelling for Archaeological Heritage Management: A research agenda*. Nederlandse Archeologische Rapporten 29, 13-23. Amersfoort: Rijksdienst voor het Oudheidkundig Bodemonderzoek

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

Leusen, P.M. van, J. Deeben, D. Hallewas, P. Zoetbrood, H. Kamermans and Ph. Verhagen 2002. *Predictive Modelling for Archaeological Heritage Management in the Netherlands. Baseline Report*. Den Haag: NWO

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

Peeters, J.H.M. 2007 *Hoge Vaart-A27 in context: towards a model of mesolithic-neolithic land use dynamics as a framework for archaeological heritage management*. Amersfoort: Rijksdienst voor Archeologie, Cultuurlandschap en Monumenten

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

Verhagen, Ph., H. Kamermans, M. van Leusen, J. Deeben, D. Hallewas and P. Zoetbrood in press. First thoughts on the incorporation of cultural variables into predictive modelling. In F. Niccolucci (ed.), *Beyond the artefact - Proceedings of CAA2004 - Prato 13-17 April 2004*. Budapest: Archaeolingua

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

Willems, W.J.H. 2007. The Times They Are A-Changin': Observations on Archaeology in a European Context. In G. Cooney (ed.), *Archaeology in Ireland: A Vision For The Future*, 5-23. Dublin: Royal Irish Academy Forum

Woodman, P.E. and M. Woodward 2002. The use and abuse of statistical methods in archaeological site location modelling. In D. Wheatley, G. Earl and S. Poppy (eds), *Contemporary Themes in Archaeological Computing.* Oxford: Oxbow Books

Appendix - Papers on predictive modelling published by members of the research team

2002

Deeben, J.H.C., D.P. Hallewas and Th.J. Maarlevelt 2002. Predictive modelling in archaeological heritage management of the Netherlands: the indicative map of archaeological values (2nd generation). *Berichten ROB* 45, 9-56. Amersfoort: Rijksdienst voor het Oudheidkundig Bodemonderzoek

Deeben, J.H.C., D.P. Hallewas and P.A.M. Zoetbrood 2002. Valuation and Selection of Late Medieval Sites in the Netherlands. In P.J. Woltering, W.J.H. Verwers and G.H. Scheepstra (eds), *Middeleeuwse toestanden. Archeologie, geschiedenis en monumentenzorg*, 451-465. Assen: Van Gorcum

Kamermans, H. 2002. The answer is blowin' in the wind. Research desires and data possibilities. In G. Burenhult and J. Arvidsson (eds*), Archaeological Informatics: Pushing the Envelope*. *CAA 2001.* BAR International Series 1016, 79-83. Oxford: Archaeopress

Leusen, P.M. van, J. Deeben, D. Hallewas, P. Zoetbrood, H. Kamermans and Ph. Verhagen 2002. *Predictive Modelling for Archaeological Heritage Management in the Netherlands*. *Baseline Report*. Den Haag: NWO

2003

Deeben, J., D. Hallewas, H. Kamermans, M. van Leusen, Ph. Verhagen, M. Wansleeben and P. Zoetbrood 2003. Predictive Modelling for Archaeological Resource Management: Development of Best Practice. In M. Doerr and A. Sarris (eds), *CAA2002. The Digital Heritage of Archaeology. Computer Applications and Quantitative Methods in Archaeology*, 430. Athens: Hellenistic Ministry of Culture

Deeben, J. and D. Hallewas 2003. Predictive Maps and Archaeolgical Heritage Management in the Netherlands. In J. Kunow and J. Müller (eds), *Landschaftsarchäologie und geographische Informationssysteme: Prognosekarten, Besiedlungsdynamik und prähistorische Raumordnungen. 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, 107-118. Wünsdorf: Brandenburgisches Landesamt für Denkmalpflege und Archäologisches Landesmuseum

Kamermans, H. 2003. Predictive Maps and Land Quality Mapping. In J. Kunow and J. Müller (eds), *Landschaftsarchäologie und geographische Informationssysteme: Prognosekarten, Besiedlungsdynamik und prähistorische Raumordnungen. 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, 151-160. Wünsdorf: Brandenburgisches Landesamt für Denkmalpflege und Archäologisches Landesmuseum

2004

Kamermans, H., J. Deeben, D. Hallewas, M. van Leusen, Ph. Verhagen and P. Zoetbrood 2004. Deconstructing the Crystal Ball: the state of the art in predictive modelling for archaeological heritage management in the Netherlands. In Stadtarchäologie Wien (ed.), *Enter the Past. The E-way into the Four Dimensions of Cultural Heritage*, BAR International Series 1227, 175 and CD-ROM (25 pages). Oxford: Archaeopress

2005

Deeben, J. and B. Groenewoudt 2005. The expanding role of predictive modeling in archaeological heritage management in the Netherlands. In C. Mathers, T. Darvill and B.J. Little (eds). *Heritage of Value, Archaeology of Renown. Reshaping Archaeological Assessment and Significance*, 298-300. Gainesville: University Press of Florida

Kamermans, H. 2005. Searching for Tools to Predict the Past; the Application of Predictive Modelling in Archaeology. In *Reading Historical Spatial Information from around the World: Studies of Culture and Civilization Based on Geographic Information Systems Data*. Proceedings of the 24th International Research Symposium. Kyoto 7-11 February 2005369-378. Kyoto: International Research Centre for Japanese Studies

Kamermans, H., J. Deeben, D. Hallewas, P. Zoetbrood, M. van Leusen and Ph. Verhagen 2005. Project proposal. In M. van Leusen and H. Kamermans (eds), *Predictive Modelling for Archaeological Heritage Management: A research agenda*. Nederlandse Archeologische Rapporten 29, 13-23. Amersfoort: Rijksdienst voor het Oudheidkundig Bodemonderzoek

Leusen, P.M. van, J. Deeben, D. Hallewas, P. Zoetbrood, H. Kamermans and Ph. Verhagen 2005. A Baseline for Predictive Modelling in the Netherlands. In M. van Leusen and H. Kamermans (eds), *Predictive Modelling for Archaeological Heritage Management: A research agenda*. Nederlandse Archeologische Rapporten 29, 25-92. Amersfoort: Rijksdienst voor het Oudheidkundig Bodemonderzoek

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

Leusen, P.M. van and H. Kamermans 2005. Introduction. In M. van Leusen and H. Kamermans (eds), *Predictive Modelling for Archaeological Heritage Management: A research agenda*. Nederlandse Archeologische Rapporten 29, 7-12. Amersfoort: Rijksdienst voor het Oudheidkundig Bodemonderzoek

Verhagen, Ph. 2005. Prospecting Strategies and Archaeological Predictive Modelling. In M. van Leusen and H. Kamermans (eds), *Predictive Modelling for Archaeological Heritage Management: A research agenda*. Nederlandse Archeologische Rapporten 29, 109-121. Amersfoort: Rijksdienst voor het Oudheidkundig Bodemonderzoek

Verhagen, Ph., J. Deeben, D. Hallewas, P. Zoetbrood, H. Kamermans and M. van Leusen 2005. A review of predictive modeling for archaeological heritage management in the Netherlands. In J.-F. Berger, F. Bertoncello, F. Braemer, G. Davtian and M. Gazenbeek (eds), *Temps et espaces de l'homme en société, analyses et modèles spatiaux en archéologie. XXVe rencontres internationales d'archéologie et d'histoire d'Antibes*, 83-92. Antibes: Éditions APDCA

2006

Kamermans, H. 2006. Problems in Paleolithic land evaluation: a cautionary tale. In M. Mehrer and K. Wescott (eds), *GIS and Archaeological Predictive Modeling*, 97-122. Boca Raton: CRC Press

Kamermans, H. 2006. Searching for Tools to Predict the Past; the Application of Predictive Modelling in Archaeology. In Uno Takao (ed.), *Reading Historical Spatial Information from around the World: Studies of Culture and Civilization Based on Geographic Information Systems Data*. February 7-11 2005, 35-46. Kyoto: International Research Centre for Japanese Studies

Verhagen, Ph. 2006. Quantifying the Qualified: the Use of Multi-Criteria Methods and Bayesian Statistics for the Development of Archaeological Predictive Models. In M. Mehrer and K. Wescott (eds), *GIS and Archaeological Predictive Modeling*, 191-216. Boca Raton: CRC Press

Verhagen, Ph., H. Kamermans and M. van Leusen 2006. Whither archaeological predictive modelling? In W. Börner and S. Uhlirz (red.). *Workshop 10. Archäologie und Computer. Kulturelles Erbe und Neue Technologien*, CD-ROM, 15 pp. Wien: Phoibos Verlag

2007

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

2008

Kamermans, H. 2008. Smashing the crystal ball. A critical evaluation of the Dutch national archaeological predictive model IKAW. *International Journal of Humanities and Arts Computing* 1 (1) 2007, 71–84. Edinburg: Edinburgh University Press

Verhagen, Ph. 2008. Testing archaeological predictive models: a rough guide. In A. Posluschny, K. Lambers and I. Herzog (eds), *Layers of Perception. Proceedings of the 35th International Conference on Computer Applications and Quantitative Methods in Archaeology (CAA), Berlin, Germany, April 2–6, 2007*. Kolloquien zur Vor- und Frühgeschichte, Vol. 10, 285-291. Bonn: Dr. Rudolf Habelt GmbH

Verhagen, Ph, H. Kamermans and M. van Leusen 2008. The future of archaeological predictive modelling. *Proceedings of Symposium The Protection and Development of the Dutch Archaeological Historical Landscape: The European Dimension, 20-23 May 2008, Lunteren*

Verhagen, Ph., M. van Leusen en H. Kamermans 2008. Een nieuwe impuls voor de archeologische verwachtingskaart. *Archeobrief* 12 (3), 27-34

In press

Kamermans, H. in press. The Application of Predictive Modelling in Archaeology: Problems and Possibilities. In F. Niccolucci (ed.), *Beyond the artefact - Proceedings of CAA2004 - Prato 13-17 April 2004*. Budapest: Archaeolingua

Verhagen, Ph., H. Kamermans, M. van Leusen, J. Deeben, D. Hallewas and P. Zoetbrood in press. First thoughts on the incorporation of cultural variables into predictive modelling. In F. Niccolucci (ed.), *Beyond the artefact - Proceedings of CAA2004 - Prato 13-17 April 2004*. Budapest: Archaeolingua

Verhagen, Ph., M. van Leusen, B. Ducke, A. Millard and H. Kamermans, in press: The bumpy road to incorporating uncertainty in predictive modelling. *Proceedings CAA2008*. Budapest: Archaeolingua

2. The future of archaeological predictive modelling[5]

Philip Verhagen[6], Hans Kamermans[7] and Martijn van Leusen[8]

2.1 Introduction

In general, academic archaeologists have been sceptical of, and sometimes even averse to, predictive modelling as practiced in archaeological heritage management (AHM) (see Van Leusen *et al*. 2005). The models produced and used in AHM are not considered sophisticated enough, and many of the methodological and theoretical problems associated with predictive modelling have not been taken aboard in AHM. At the same time, the production and use of predictive models has become a standard procedure in Dutch AHM (Deeben *et al*. 1997; 2002; Deeben 2008), and it has clearly attracted interest in other countries as well.

The main reason for using predictive models in AHM is efficiency. In 'post-Malta' archaeology, the financial, human and technical resources allocated to archaeology have increased enormously. But at the same time, these resources have to be spent both effectively and efficiently. So why not create and use tools that will allow us to do so? Archaeological predictive models will tell us where we have the best chances of encountering archaeology. Searching for archaeology in the high probability areas will 'pay off', as more archaeology will be found there than in low probability zones. It is a matter of priorities: we can not survey everything, and we do not want to spend money and energy on finding nothing. And there is also the political dimension: the general public wants something in return for the taxpayers' money invested in archaeology. It's not much use telling politicians to spend money on research that will not deliver an 'archaeological return'. But 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? These are hard choices indeed for those involved in AHM. Archaeologists who do not have to make these choices have an easy job: they can criticize the current approach to predictive modelling from the sidelines, and do not have to come up with an alternative.

Within the BBO program we have been trying to provide such an alternative to the archaeological community (Kamermans *et al*. 2005). However, at the end of the project, we have to conclude that we have only been partly successful. We have done a fair amount of research, published three books and many papers, made the problems with predictive modelling internationally visible but failed to change the procedures of predictive modelling in the Netherlands. In this paper we venture to offer some explanations for the lack of success of new approaches to predictive modelling in AHM in the Netherlands up to now. And finally, we will try to sketch the future of archaeological predictive modelling, for which we can see three distinct scenarios.

2.2 Problems and solutions

Over the past twenty-five years, archaeological predictive modelling has been debated within the larger context of GIS applications in archaeology, and the processual/post-processual controversy that has dominated the archaeological theoretical debate since the late 1980s (see Van Leusen *et al*. 2005). Predictive modelling is rooted in the processual tradition, with its emphasis on generalization and quantitative 'objective' methods, and its lack of interest in the subjective and individual dimensions of archaeology. In itself, this is not a matter of 'bad' versus 'good' archaeology. Within the context of AHM, generalized maps are necessary tools to reduce the enormous complexity of archaeology to manageable proportions. However, the lack of real interest in using spatial technology and statistical methods in post-processual academic archaeology has certainly slowed down the development of predictive modelling as a scientific method. The feeling that processual approaches no longer offered a real contribution to the advancement of archaeological science has left predictive modelling

[5] A slightly different version of this chapter has been published as Verhagen *et al*. 2008.

[6] ACVU-HBS, Amsterdam, the Netherlands.

[7] Faculty of Archaeology, Leiden University, the Netherlands.

[8] GIA, Groningen University, the Netherlands.

somewhat 'lost in space' it seems. Which is a pity, because even if we do not want to use predictive modelling in an AHM context, there still is a lot of potential in spatial technologies ('GIS') to develop and test theories of spatial patterning of settlements and human activities.

The criticism of predictive modelling in scientific literature has focused on three main issues: statistics, theory and data. In all three areas, predictive modelling as it stands today is considered by various authors to insufficiently address the complexity of the matter. Statistical methods are used uncritically, often using a limited number of techniques that are not the best suited around. Archaeological theory, especially where it concerns the human and temporal factors in site placement, only plays a marginal role in selecting the variables used for predictive modelling. And archaeological data, which we all know have various degrees of reliability, are used without much source criticism. And while this is all very much true, and many archaeological predictive maps are rather coarse representations of a complex archaeological reality, these criticisms mask a more fundamental question: what is the quality that is *needed* of a predictive model? Because this is precisely why models are made that are not very sophisticated from a scientific point of view: they are considered good enough for the purposes they are made for[9].

Our one fundamental problem with predictive modelling is therefore this issue of quality. No one seems to know what constitutes a 'good' model, and no tools are available and used to make explicit the quality of the models. Within our research project, we have tried to focus on these issues by looking at the potential of new statistical techniques for incorporating uncertainty in the predictions, and by studying the best ways of testing the models (this volume chapters 7 and 8). Our first foray into the uncharted waters of model quality concerned the role of expert judgement in a quantitative framework. When the first criticisms of predictive modelling appeared in the late 1980s, it quickly became clear that a fully 'inductive' approach was in many cases unsatisfactory. The lack of reliable survey data in many areas of the world basically ruled out a rigorous statistical approach, unless vast amounts of money were invested in survey. The pragmatic solution therefore was to stop using statistical methods for developing predictive models, but instead rely on expert judgement, and see if the experts' views were corroborated by the available archaeological data. However, in doing so, a major advantage of statistical methods was thrown overboard: the ability to come up with estimates of for instance site density, in real numbers, and the calculation of confidence intervals around the estimates. Expert judgement models only classify the landscape into zones of low, medium and high probability without specifying the numbers involved. How many archaeological sites can we expect in a high probability zone? And how certain can we be of this estimate with the available data? Statistical methods will provide these numbers, expert judgement will not.

It turns out that Bayesian statistical techniques are very well suited to provide numerical estimates and confidence intervals on the basis of both expert judgement and data (Millard 2005; Verhagen 2006; Finke *et al*. 2008). However, they have not been used in anger in predictive modelling so far. This is probably because of the relative complexity of the calculations involved. There are very few archaeologists in the world that can perform these calculations, even though computing power is now no longer an obstacle. We have however proved that it can be done, and we see Bayesian statistics as a very powerful and useful tool for predictive model building (this volume chapter 9).

We also tested the potential of Dempster-Shafer modelling (this volume chapter 9), which has been suggested as an alternative to 'standard' statistical methods as well (Ejstrud 2003; 2005). While the results of earlier modelling studies indicated that it performed better than most statistical tools including Bayesian analysis it is only partly suitable to include expert judgement. Furthermore, its conceptual basis is rather complex. We will not go into detail here; suffice it to say that Dempster-Shafer modelling is more controversial in statistical science than Bayesian statistics, and it is more difficult to understand.

However, even if tools like Bayesian statistics can build a bridge between expert judgement and quantification, we still need reliable data to have it delivering its potential. The testing issue is therefore of

[9] However not all users consider the maps good enough (see this volume chapter 6).

primary importance to predictive modelling. What is probably most problematic in this respect is the lack of attention by archaeologists to the simple statistical principle of data representativity. Nobody seems to wonder if the available archaeological sample represents in any way the original archaeological population. No matter what statistical method is used, this issue needs to be addressed first before attempting to produce a numerical estimate of any kind. And while it is possible to reduce the bias encountered in existing archaeological survey data to an acceptable level, in order to have reliable archaeological predictive models we also need to survey the low probability zones. So here we are facing a real paradox: predictive models are developed to reduce the amount of survey (or even skip it) in low probability zones, yet statistical rigour tells us to do survey there as well.

Our approach has been to re-assess the value of using statistical methods in predictive modelling (this volume chapters 7 and 8). We are convinced that this re-assessment is necessary, and think that it can offer a valuable contribution to AHM as well. If we can base the models on sophisticated statistical methods and reliable data, then we can really start using predictive models as archaeological and/or economic risk management tools.

In the end, we have not been able to get this message across to the AHM community in the Netherlands yet. While we have not done an exhaustive survey among our colleagues, we think that the following reasons may be responsible for it:

1. the innovations suggested are too complex. While it is sometimes said that statistics are not very difficult, but only very subtle, in practice most archaeologists are not working with it on a daily basis. Some even have difficulty grasping the most fundamental principles of quantitative methods. This makes it hard to get the message across, as it does not really help when we have to bridge a large gap in knowledge between the statistical experts and the people who have to use the end results of statistical models.
2. shifting from the current expert judgement approach to a more sophisticated statistical approach is too expensive. Improving the models in the way we suggest does not *replace* anything in the current way of dealing with predictive modelling, it only *adds* to it. So, on top of the things we already do, like gathering and digitising all the available information, and interviewing the experts, we now also need to have a statistical expert doing the modelling, a data analysis programme to detect and reduce survey bias, and perhaps even a test survey as well.
3. it is irrelevant. While academic researchers may be bothered with the quality of the models, most end users are not. They trust the experts. Especially those responsible for political decision making will not care, as they only need clear lines drawn on a map, that will tell them where to do survey and where not. If the archaeologists are happy with it, then they are as well.
4. and this ties in to our last explanation: archaeologists may have reason to be afraid for more transparent methods that will give insight into the uncertainties of predictive models to non-archaeologists. When anyone can judge model quality, they will lose their position of power in dealing with politicians and developers.

2.3 Recommendations

For the moment we have not been able to convince our AHM colleagues of the need to improve predictive modelling along the lines suggested. We certainly did not have enough time and money to fully develop these new approaches into practical working solutions, but at least we have showed some promising avenues to follow. From our research, a number of recommendations result for each of the themes originally identified in the baseline report (van Leusen *et al*. 2005).

Concerning the quality and quantity of archaeological input data we recommend:

- to use sophisticated statistical methods (Bayesian statistics, Dempster-Shafer modelling) to integrate expert judgement and archaeological data in one quantitative framework
- to improve the way of registering archaeological fieldwork data, including the methods of data gathering, and the location of areas where no archaeological remains were found
- to analyse archaeological fieldwork data for the presence of systematic biases that are either the consequence of data collection strategies, or of geological conditions, and use methods for correcting these biases to come up with more representative data sets

Concerning the relevance of environmental input data we recommend:

- to use new relevant data sources like LIDAR elevation models, geological borehole databases, historical maps and remote sensing images
- to explicitly describe and judge the quality and reliability of environmental input data

Concerning the need to incorporate social and cultural input data we recommend:

- to take a critical look at modern archaeological theory, and see where the social and cultural factors involved in site location can be translated into spatial predictive models

Concerning the lack of temporal and/or spatial resolution we recommend:

- to always produce 'diachronic' predictive models, in which both environmental and cultural dynamics over the long term are represented
- to adapt the scale of the mapping to the intended use; maps that are only meant to provide a baseline for spatial planning zoning can be less detailed than maps that should also provide detailed advice on the type of archaeological research that should be done in those zones

Concerning the use of spatial statistics we recommend:

- to apply appropriate statistical methods for modelling (like Dempster-Shafer modelling) and testing (like resampling), in order to prevent creating spurious correlations and over-optimistic views of model quality

Concerning the testing of predictive models we recommend:

- to set up long-term field testing programs, focusing on the areas where uncertainties are highest
- to use appropriate statistical methods to judge the quality of predictive models

Most of these recommendations will not come as a surprise, even though we place a stronger emphasis on statistical rigour than is currently usual. Improvements in the quality of the environmental input data and the temporal and spatial resolution of predictive modelling are already clearly observed over the past few years. Even the inclusion of social and cultural factors in predictive models is now gradually becoming a feasible option, and can easily fit into academic research projects. In contrast, the rigorous testing of predictive models is still far from becoming a standard procedure. The reasons for this are on the one hand found in the highly scattered nature of 'post-Malta' archaeological prospection that is only done in high-probability areas were survey is enforced. Field testing, while sometimes part of a predictive modelling project, is usually only done to check the modelling assumptions: are *e.g.* the geomorphological units mapped correctly, and where are the highly disturbed zones? Systematic collection and analysis of survey results is a time-consuming task, and when it is done, it is primarily used to produce up-to-date catalogues of archaeological finds, rather than for purposes of model development and testing – which in many cases would imply collecting data for larger areas than the region modelled. On the other hand we observe a lack of interest in the issue of model quality by authorities from the national to the local level. This may partly be attributed to the fact that archaeology is not such a big economic risk after all; developers and authorities do not expect a direct benefit from investing in risk management methods. The *scientific* risk involved however should worry all archaeologists.

There is no way to prevent the development of a highly biased archaeological record if we do not take the issue of predictive model quality and testing seriously. This is primarily a task for the archaeological community itself, including companies, universities and the national archaeological service. And yes, it will

need money – which implies defining different priorities and making different choices when applying for funding and when spending budgets. The current procedures in Dutch archaeological heritage management however preclude prioritising this kind of research, as they are purely focused on selecting and eventually excavating individual archaeological sites. The recently published National Archaeological Research Agenda will probably only reinforce the focus on localized interpretive archaeological research, rather than on methodological improvement and synthetic research.

2.4 Conclusions

What will be the future of predictive modelling in AHM in the Netherlands? We think that there are three scenarios that may be followed in the next ten years or so. The first one is a scenario of 'business as usual'. There are some points that speak in favour of this option. First of all, predictive maps are fully accepted by the Dutch archaeological community and are relatively well embedded in planning procedures. Secondly, as far as is known, the use of the predictive maps that are currently around has not led to archaeological disasters, even though some grumbling is occasionally heard about the quality of the maps. Rather, it is the other way around. Predictive maps, in spite of there theoretical and methodological shortcomings, are more than effective in protecting the archaeological heritage: they over-protect. This of course is an unwanted situation. In practice, municipal authorities commissioning predictive maps for their own territory do this with the explicit aim of reducing the area where preliminary research is needed. We might be heading to a future where commercial predictive modelling will have as its highest aim the reduction of the zones of high archaeological probability – without having the tools to judge whether this reduction is supported by the archaeological data.

Cautious archaeologists would therefore certainly prefer the second possible scenario: this is, to stop using predictive models, and do a full survey of all the threatened areas. Obviously there are many countries in the world that can do archaeological heritage management without predictive maps. Even in the United States, full survey is sometimes seen as a feasible alternative. In its favour speaks the reassuring thought that all archaeological remains present will be detected, and if necessary protected or excavated. However, this scenario is a political impossibility in the Netherlands, for the reasons mentioned above: politicians want less money to be spent on archaeology, not more. And even in countries where full survey is supposedly done, as in France or the United Kingdom, preliminary selection by means of desk-based assessment plays an important role in deciding where to do what kind of archaeological research. The question then is: is the Dutch method of selection by means of predictive maps better than doing the desk-based assessments that are common practice in France and the United Kingdom?

There is no way we can escape doing selections in archaeological heritage management. However, we need adequate tools on which to base these selections. Which brings us to the third scenario: the further development of predictive models into true risk assessment tools? There are, we feel, at least three supporting arguments for moving towards quantitative mapping of model uncertainty. First of all, there is the question of model quality and testing we already discussed. At the moment, expert judgment is among the inputs used to determine whether a zone is placed into high, medium of low probability, and uncertainties regarding this classification are never specified. However, expert judgment can never serve as an independent criterion of model quality. For independent model testing, we need data sets based on representative samples of the archaeological site types predicted. Secondly, the absence of estimates of the uncertainties in predictive models may lead to 'writing off' zones of low probability that are in fact zones where little archaeological research has been done. By including uncertainty measures in the models, it may be possible to break through the vicious circle of self-fulfilling prophecies that is created by doing ever more surveys in zones of high probability. And thirdly, the use of true statistical estimates and confidence intervals brings with it the perspective of making risk assessments in euros, rather than in relative qualifications of site density. Predictive modelling then may provide a first assessment of the bandwidth of the archaeological costs of a development plan.

References

Deeben, J., D. Hallewas, J. Kolen, and R. Wiemer 1997. Beyond the crystal ball: predictive modelling as a tool in archaeological heritage management and occupation history. In W. Willems, H. Kars, and D. Hallewas (eds), *Archaeological Heritage Management in the Netherlands. Fifty Years State Service for Archaeological Investigations*, 76-118. Amersfoort: Rijksdienst voor het Oudheidkundig Bodemonderzoek

Deeben, J.H.C., D.P. Hallewas and Th.J. Maarlevelt 2002. Predictive modelling in archaeological heritage management of the Netherlands: the indicative map of archaeological values (2nd generation). *Berichten ROB* 45, 9-56. Amersfoort: Rijksdienst voor het Oudheidkundig Bodemonderzoek

Deeben, J.H.C. (ed.) 2008. *De Indicatieve Kaart van Archeologische Waarden, derde generatie*. Rapportage Archeologische Monumentenzorg 155. Amersfoort: RACM

Ejstrud, B. 2003. Indicative Models in Landscape Management: Testing the Methods. In J. Kunow and J. Müller (eds), *Landschaftsarchäologie und geographische Informationssysteme: Prognosekarten, Besiedlungsdynamik und prähistorische Raumordnungen. 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, 119-134. Wünsdorf: Brandenburgisches Landesamt für Denkmalpflege und Archäologisches Landesmuseum

Ejstrud, B. 2005. Taphonomic Models: Using Dempster-Shafer theory to assess the quality of archaeological data and indicative models. In M. van Leusen and H. Kamermans (eds), *Predictive Modelling for Archaeological Heritage Management: A research agenda*. Nederlandse Archeologische Rapporten 29, 183-194. Amersfoort: Rijksdienst voor het Oudheidkundig Bodemonderzoek

Finke, P.A., E. Meylemans and J. Van de Wauw 2008. Mapping the Possible Occurrence of Archaeological Sites by Bayesian Inference. *Journal of Archaeological Science* 35, 2786-2796

Kamermans, H., J. Deeben, D. Hallewas, P. Zoetbrood, M. van Leusen and Ph. Verhagen 2005. Project proposal. In M. van Leusen and H. Kamermans (eds), *Predictive Modelling for Archaeological Heritage Management: A research agenda*. Nederlandse Archeologische Rapporten 29, 13-23. Amersfoort: Rijksdienst voor het Oudheidkundig Bodemonderzoek

Leusen, P.M. van, J. Deeben, D. Hallewas, P. Zoetbrood, H. Kamermans and Ph. Verhagen 2005. A Baseline for Predictive Modelling in the Netherlands. In M. van Leusen and H. Kamermans (eds), *Predictive Modelling for Archaeological Heritage Management: A research agenda*. Nederlandse Archeologische Rapporten 29, 25-92. Amersfoort: Rijksdienst voor het Oudheidkundig Bodemonderzoek

Millard, A. 2005. What Can Bayesian Statistics Do For Predictive Modelling? In M. van Leusen and H. Kamermans (eds), *Predictive Modelling for Archaeological Heritage Management: A research agenda*. Nederlandse Archeologische Rapporten 29, 169-182. Amersfoort: Rijksdienst voor het Oudheidkundig Bodemonderzoek

Verhagen, Ph. 2006. Quantifying the Qualified: the Use of Multi-Criteria Methods and Bayesian Statistics for the Development of Archaeological Predictive Models. In M. Mehrer and K. Wescott (eds), *GIS and Archaeological Predictive Modeling*, 191-216. Boca Raton: CRC Press

Verhagen, Ph, H. Kamermans and M. van Leusen 2008. The future of archaeological predictive modelling. *Proceedings of Symposium The Protection and Development of the Dutch Archaeological Historical Landscape: The European Dimension, 20-23 May 2008, Lunteren*

3. On costs and benefits in archaeological prospection

Marten Verbruggen[10]

3.1 Introduction

in Dutch archaeological heritage management, an important role is given to prospective research. Due to the implementation in Dutch legislation of the Treaty of Malta, project developers are obliged to have archaeological sites mapped at their cost. Cost-benefit analysis, with the question whether the financial efforts weigh up against the archaeological results, has thus entered archaeology. After a first success of the small-scale and broad inductive predictive maps, the demand is increasing sharply for large-scale, detailed deductive maps. Municipalities are the main commissioners of these maps as they have been given far-reaching powers in the field of the preservation of monuments by the recent amendment of the law. It looks as if the success of the inductive predictive maps has to come to an end in favour of large-scale deductive maps.

3.2 Costs and benefits in archaeological research

Anybody involved in scientific research knows that choices have to be made continuously. As the resources for research are always restricted, every researcher will have to limit himself in the choice of problem statements and research methods to ensure that the results will be available within the agreed time. Choices will not only have to be made at the start of the research but also during the execution, as research rarely runs exactly as thought or hoped. The choices in principle deal with questions *within* the research and questions *about* the research. The questions within the research, centre on which research method is appropriate and attainable, how large a sample should be, how the reliability can be increased, etc. The questions about the research are usually limited to the key question whether all the efforts will make up for the expected results. In fact this is a cost-benefit analysis that should be completed favourably prior to the research. But also during the research this question can be topical, for instance when the interim results are disappointing or when the efforts are much greater than assumed earlier.

When the researcher has been given a grant for his research, he (or she) is responsible for the budget and in principle he is 'his own master' over his research. He is autonomous in making choices in his 'own' research, based on scientific considerations. The only obligation he has is to account for his choices in the research report and to underpin these scientifically.

From the moment that the archaeologist carries out developer-funded research, he is no longer master of the research. This situation often prevails in Dutch archaeological heritage management, where the project developer is obliged to have prospective research carried out in order to map archaeological values. The project developer bears the costs of the research and the government decides the requirements to be fulfilled by the research. The researcher is required to adhere to certain quality requirements. In practice this means following the procedures and specifications of the Dutch Archaeology Quality Standard (Kwaliteitsnorm Nederlandse Archeologie, usually referred to as the KNA, version 3.1; SIKB 2006).

In prospective research in Dutch archaeology, there are thus generally three parties involved, the interests of whom could be incompatible. The interest of the government is to make a well-founded (selection) decision based on research regarding the sites found; the researcher wants above all that his research meets all scientific requirements, and the project developer wants to 'buy in' research at as low a price as possible, and of a standard that will be fully acceptable to the government. As a consequence, the researcher will have to justify all choices in his research. Crucial here is a balancing of costs and benefits. Archaeological predictive maps can play a prominent role here.

[10] RAAP Archeologisch Adviesbureau (RAAP archaeological consultancy). Amsterdam, the Netherlands.

3.3 Predictive modelling

The introduction of *predictive modelling* and more specifically of predictive maps is in the Netherlands closely linked to the discussion on costs and benefits in prospective research. The 'product' predictive map is relatively new in Dutch archaeology. The first map was published only in 1990 – it concerned a so-called potential map of the Rijssen-Wierden area – (Ankum and Groenewoudt 1990). At that time, many large-scale land consolidations were carried out and there was a risk that many archaeological sites would be destroyed by earthmovers. To prevent this from happening, the whole surface of thousands of hectares would have to be investigated archaeologically, with resulting high costs. As the Government Service for Land Management (Landinrichtingsdienst) would bear 90% of the costs, it came as no surprise that this same service became the commissioner of the map mentioned earlier. The main aim was to divide the area into zones of high and low archaeological potential. Subsequently, field research would only be carried out in the zones of high potential. Clearly the aim was a sound deployment of scarce resources. At the end of the summary of the accompanying report, the motivation of the research is plainly stated: "Analysis of location factors and the use of predictive models can increase the yield of archaeological inventories". We are clearly dealing here with a cost-benefit analysis and that is a very legitimate aim. But what exactly is 'the return' of a predictive map in prospective research? Does the emphasis lay here solely on the number of archaeological sites or on the quality thereof?

Since the initial research of RAAP, many maps have been published, special mention deserve the Indicative Map of Archaeological Values (Indicatieve Kaart van Archeologische Waarden, IKAW first, second and third generation, Deeben *et al*. 1997; 2002; Deeben 2008), the series of provincial cultural-historical value maps, and dozens of local authority predictive maps (cf. Jansen & Roymans 2002; Lotte and Tebbens (red) 2005; Van Wijk and Van Hoof 2005). In addition, a steady flow of scientific publications on *predictive modelling* started up (see Deeben *et al*. 2002). Hence predictive maps play a role in Dutch archaeological heritage management that cannot be overlooked. In the course of the more then 15 years that these maps have been around, it has however become clear that the aim with which these maps were made and the role they subsequently played in the decision-making process in the archaeological heritage management have changed radically.

3.3.1 Inductive predictive maps

At the time of the introduction of inductive predictive maps, the State and the provinces mainly decided the policy of archaeological monument preservation. From their responsibility arose a strong desire for a reliable picture of the existing and to-be-expected archaeological sites. Due to the availability of an extensive dataset of archaeological find spots (the national database for archaeological find spots ARCHIS) and qualitatively good soil maps, small-scale inductive predictive maps of the entire Dutch territory could be made quickly and simply with the use of a Geographic Information System (GIS). Thus between 1996 and 2002 two versions of the IKAW appeared (Deeben *et al*. 1997; 2002), and in its wake a series of provincial predictive maps. These maps were very successful, as for policy makers at state and provincial level they offered exactly the information required for their decision-making process; the question in which areas prospective research should or should not be carried out was generally the most important reason for consulting the IKAW. The scale, and the simple and orderly legend structure of the IKAW were clear advantages for this type of user. With one glance at the map it could be decided in which areas the chance of the presence of archaeological sites was high (red), low (yellow), or intermediate (orange). In the weighing of where prospective research should be carried out, the legend colours soon became guiding: in the (red) areas with high expected densities research was carried out, and in orange areas building activities had to be supervised archaeologically. Yellow areas were released for development without any form of research.

The scientifically minded archaeologists were less happy with these inductive maps (see Van Leusen and Kamermans 2005). Their main objection was that the dataset used for the production of the IKAW was heavily distorted by an overrepresentation of sites on or close to the surface. These could be found in particular in the Pleistocene part of the country, where characteristically many of these sites had been heavily affected by cultivation or other disturbances. Sites in the Holocene part of the country were strongly underrepresented

due to their poor visibility as a result of covering sediments. Based on palaeogeographical research and some chance finds, it had in the meantime become known that many sites should be present in the Holocene part of the country, with the advantage that these sites would undoubtedly be very complete due to their location several metres deep under the surface. Unfortunately they were in the yellow IKAW areas where research generally was not enforced by the higher authorities.

In daily practice, areas with a low indicative value were not investigated further and areas with a high one were. The predictive maps of the RACM and provinces (also with a small map scale) were used as 'decision rule', with a very restricted interpretation of the notion of yield: the expected number of find spots were guiding in the choice whether or not prospective research was required. The result was that in the past years many find spots worthy of preservation have been lost because they 'were unlucky enough' to lie in zones with a low indicative value. Instead of increasing the 'archaeological return', the predictive map had soon become a means to save costs. And in the end it even decreases the archaeological return, if we may assume that in areas with a high indicative value, the number of (partly disturbed) find spots not worthy of preservation is relatively high[11].

However, the success of the inductive predictive maps seems to come to an end due to the recent change in the monuments legislation, in favour of deductive predictive maps, where expert judgement has been used.

3.3.2 Deductive predictive maps

Coinciding with the appearance of the IKAW and the provincial predictive maps, the contours became visible of the new legislation on archaeological monument preservation in the Netherlands. The most important changes were decentralization of policy, and privatisation of archaeological research that in the past was reserved for the authorities and universities. Two new groups of commissioners arose. The first group consisted of the nearly 450 municipalities in the Netherlands. These took over the authority from the State and provinces regarding archaeological monument preservation and the concomitant research. Municipalities required in particular detailed maps for their relatively small territory. With it, the interest in 'densities of sites' shifted to 'quality of sites'. This was because municipalities were not motivated to enforce a research obligation in the complex process of weighing of interests if there is little chance of sites worthy of preservation. Hence, interest went increasingly to the sites that were in the orange and yellow zones on the inductive maps.

The second group consisted of a motley collection of so-called 'developers'. That is to say, anybody initiating a soil disturbance: project developers, farmers, small private individuals, etc. Also this group had a need in particular for detailed maps for their planning, something in which inductive maps could not oblige.

The combination of wishes from municipalities and developers resulted in a steady stream of deductive predictive maps with the advantages of great detail and attention to the quality of sites instead of quantity. The detailed maps provided yet another advantage that was connected with the stepped process of archaeological prospection. It became possible to carry out again and again a cost-benefit analysis during the research process. These analyses do however have the consequence that the results of inventorizing field research cannot be used directly as input in inductive predictive models. I will come back to this at the end of this paper.

3.4 The dutch archaeology quality standard (Kwaliteitsnorm Nederlandse Archeologie, KNA)

In 2001, and in advance of the new legislation, the Dutch Archaeology Quality Standard (Kwaliteitsnorm Nederlandse Archeologie, KNA) was introduced (Voorbereidingscommissie Kwaliteitszorg Archeologie 2001). The backbone of this standard was the detailed description of the process of archaeological monument preservation in several protocols. Thus the research prior to the decision to excavate, protect or release the planned area was split up into the protocols of desk research and inventorizing field research. Within these protocols several prospection techniques were described, with the requirement that the selection of a technique

[11] The proportion of find spots not worthy of preservation will be more or less the same in areas with a low, medium or high indicative value, but the absolute number of find spots will be higher in an area with a high density of find spots (HK).

should be substantiated on the basis of a specified archaeological expectation that was the result of the desk research. In the specified expectation the aim was explicitly not at the different expected *densities* of sites, but at the *prospection characteristics* thereof. These characteristics (size, depth location, type and density of the expected finds) after all determined the choice between the various prospection techniques, such as surface mapping, trial trenches or augering research. Apart from a subdivision in techniques, a distinction was also made into an exploratory, a mapping and an evaluation phase. The exploratory phase was meant to gain more insight into the geology of the planned area, with the aim of making a distinction between areas with a high chance and a low chance on archaeological finds. In a way, the result of this phase is a (deductive) predictive map, with the characteristic that it concerns a relatively small area and that the degree of detail is large compared to the traditional inductive maps. The mapping phase was meant to locate the findspots, and the evaluating phase to evaluate the located find spots. This splitting up into phases made it possible to evaluate the previous phase time and again and to decide whether a continuation was desirable. It became thus possible to carry out a cost-benefit analysis at several instances in the archaeological process, with the key question: continue or stop.
For this analysis the following notions were introduced:

- *necessity.* At the base of every research in the Dutch archaeological heritage management is the question whether there is an obligation for the developer (the disturber) to have research carried out. The necessity arises from legislation and regulation. From the principle of necessity follows also that the research may not be more extensive than is legitimised.
- *effectiveness.* A prospection method ought to do what it promises to do, *i.e.* that the intended aim – the tracing of archaeological sites – can be achieved with the means.
- *proportionality.* A universal point of departure is that the aim is a proper balance between effort and expected results. This dilemma was discussed at the beginning of this chapter. Here the costs oppose the scientific yields. These advantages and disadvantages are however distributed unequally in archaeological heritage management – and in the eye of the disturber unfairly. However it is not possible to set rules in advance for what will be seen as proportional or disproportional by all parties. Yet, apparently "we" always seem to work it out through negotiation and we may trust to find the middle ground by the effect of precedent.
- *subsidiarity.* The principle means that always the least severe (but effective) method should be chosen. An example: if a percentage of cover of 5% for trial trenches suffices, it cannot be defended to opt for a percentage of 10%.

In daily practice the above notions are used as balancing framework for the beginning of the research and during decision moments between the exploratory and mapping phases, and between the mapping and evaluation phases. Again and again the question is at the forefront whether (subsequent) research is worth the effort and which research method should be chosen. Once a research has started, it does not in any way mean that it will actually be finished. After every phase the decision can be made to stop. One element that will play an increasingly important role in the balancing framework in the coming years is the question whether it is desirable to trace *all* archaeological sites expected in the planned area, or that there is a preference for a selection thereof. The difference in costs between tracing large or very small sites is such that a cost-benefit analysis can also be carried out here. The consequence of the above is that the result of an inventorizing field research cannot simply be interpreted as presence or absence of sites. It is always:

- a result of an earlier policy choice (on the basis of for instance a predictive map) regarding the necessity of research, linked to the question of which types of sites should be traced, and
- the result of a number of decision moments in which the notions of effectiveness, proportionality and subsidiarity are decisive.

The result of carrying out the above weighing is that the outcome of an inventorizing field research cannot be used without restrictions as input for a statistical model for the purpose of predictive modelling.

3.5 Conclusions

Predictive modelling is an important tool in Dutch archaeological heritage management. It is used in the first phase of research as a tool for prospection but also for selection. As a consequence of the revised Monument and Historic Buildings Act the small-scale inductive maps are no longer relevant. The now popular large-scale deductive maps pay more attention to the quality of sites then of the quantity. These detailed maps can also be used for a cost-benefit analysis. The consequence of the fact that predictive maps play a role in the process of selection is that after the phase of inventorizing field research we still do not have a complete picture of the present or absence and the distribution of sites in a region.

References

Ankum, L.A. and Groenewoudt, B.J. 1990. *De situering van archeologische vindplaatsen*. RAAP-rapport 42. Amsterdam: Stchting RAAP

Deeben, J., D. Hallewas, J. Kolen, and R. Wiemer 1997. Beyond the crystal ball: predictive modelling as a tool in archaeological heritage management and occupation history. In W. Willems, H. Kars, and D. Hallewas (eds), *Archaeological Heritage Management in the Netherlands. Fifty Years State Service for Archaeological Investigations*, 76-118. Amersfoort: Rijksdienst voor het Oudheidkundig Bodemonderzoek

Deeben, J.H.C., D.P. Hallewas and Th.J. Maarlevelt 2002. Predictive modelling in archaeological heritage management of the Netherlands: the indicative map of archaeological values (2nd generation). *Berichten ROB* 45, 9-56. Amersfoort: Rijksdienst voor het Oudheidkundig Bodemonderzoek

Deeben, J.H.C. (ed.) 2008. *De Indicatieve Kaart van Archeologische Waarden, derde generatie*. Rapportage Archeologische Monumentenzorg 155. Amersfoort: RACM

Jansen, B. and J.A.M. Roymans 2002. *Het Land van Cuijk, gemeente Cuijk; een archeologische verwachtingskaart*. Amsterdam: RAAP-rapport 828

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

Lotte R.M.and L.A. Tebbens (eds) 2005. *Gemeente Enschede Archeologische verwachtingskaart.* Enschede / 's-Hertogenbosch: BAAC - rapport 04.238

SIKB 2006. *Kwaliteitsnorm Nederlandse Archeologie (KNA). Ontwerp herziening versie 3.1.* Gouda: SIKB

Voorbereidingscommissie Kwaliteitszorg Archeologie 2001. *Kwaliteitsnorm Nederlandse Archeologie. Eindrapport van de Voorbereidingscommissie Kwaliteitszorg Archeologie.* Den Haag: Ministerie van Onderwijs, Cultuur en Wetenschappen

Wijk, I.M. van and L.G.L. van Hoof 2005. *Stein, een gemeente vol oudheden een archeologisch beleidskaart voor de gemeente Stein*. Leiden: Archol rapport 29

4. The high price or the first prize for the archaeological predictive model

Martin Meffert[12]

4.1 Summary

The Dutch national archaeological knowledge centre, the National Service for Archaeology, Cultural Landscape and Built Heritage (RACM) has formulated an archaeological predictive model: the Indicative Map of Archaeological Values (Indicatieve Kaart Archeologische Waarden, IKAW, Deeben *et al*. 1997; 2002; Deeben 2008). In a short time, this map has had far-reaching consequences for spatial planning in the Netherlands. At the beginning of the 21st century, a large number of provinces in the Netherlands legally established this map for their territory in the framework of the Spatial Planning Act (Wet Ruimtelijke Ordening). The IKAW has since been applied as policy instrument in spatial planning. With this, 'archaeological boundaries' in the Netherlands have been fixed spatially for the first time in history, boundaries with unknown archaeological values. For an developer drawing up spatial plans, these boundaries make the difference between archaeological research or not, and between archaeological costs or not.

As is the case in all scientific models, also this model will have to be continually improved. The boundaries between areas with a high, medium or low indicative values for archaeological remains will therefore shift. These shifts will have as a consequence an increase or decrease in the scale of the archaeological research, and thus in the number of commissions in the archaeological market, and will therefore influence the archaeological employment situation. At the same time these boundary shifts will have an influence on the − higher or lower – costs in the realization of building projects.

Outside the archaeological world, these archaeological boundaries are, for the time being, seen as static. From the moment this is no longer the case, the instrument of the predictive map will be brought into action as a financial steering instrument. Those agencies with the opportunity to make archaeological predictive models, have therefore a powerful and precious weapon in their hands. And who in the Netherlands is actually not authorized to produce a predictive map? Regulatory measures have become necessary, also because in the new Monument Act the legislator did not define the notion of 'expected archaeological values' or 'expected archaeological monuments'. However the legislator did make them something worth preserving and/or obligatory for further research. This chapter will discuss the possibilities and impossibilities regarding regulatory measures concerning the production of archaeological predictive maps.

4.2 Introduction

At the beginning of this century, the Dutch government has created an archaeological market for her territory. In the Dutch system of the archaeological heritage management almost all archaeological research is carried out by commercial companies or by independent non-profit institutions. But also these institutions will have to keep their head above water on the archaeological market without public money. The only exceptions to this are the small inner-city excavations in large towns with a town archaeologist, and excavations carried out by universities as part of training. The national archaeological predictive model Indicative Map of Archaeological Values (IKAW) has played a large role as policy instrument in the creation of this archaeological market. Basically it indicates where it is unreasonable (areas with a low probability of finding archaeological remains), but especially where it is reasonable (areas with a high or medium-high probability of finding archaeological remains) to make archaeological investigation compulsory in the Netherlands.

[12] Provincial archaeologist of Noord-Brabant, the Netherlands.

The IKAW is the best-known archaeological predictive model in the Netherlands and was designed by the archaeological knowledge centre of the state, the RACM (Deeben *et al*. 1997). At present, the authorities are mainly using version 2 of this predictive model (Deeben *et al*. 2002). Recently a third version has been published (Deeben 2008).

In a short time, the IKAW has had far-reaching consequences for spatial planning in the Netherlands. At the beginning of the 21st century, a large number of provinces legally established this map for their territory in the framework of the Spatial Planning Act. This has often taken place in the framework of their spatial- or regional plans (Provinciale Staten van Noord-Brabant 2002, 64-66). The predictive map has since been applied as policy instrument. The provinces have hereby for the first time in the history of the Netherlands spatially fixed the boundaries of areas with unknown archaeological values. This will have far-reaching spatial and thus financial consequences, not only for society as a whole, but also for the private initiators.

The Netherlands has been divided into areas with four different indications for finding archaeological material: areas with a high indicative value[13], with a medium indicative value, with a low indicative value, and areas with a very low indicative value. The boundaries between these areas make the difference between archaeological costs or not for an developer when producing spatial plans. These boundaries make the difference of a compulsory archaeological preliminary investigation or not, prior to the submission of a zoning scheme with the provincial authorities, in which a request is made to change to a certain use of an area (Van Leusen *et al*. 2005, 48-49; Provinciale Staten van Noord-Brabant 2002, 64-66). For instance before agricultural land can be transformed into a new housing area or into a golf course, the aspect of archaeology will first have to be mapped and valued. In the framework of the Spatial Planning Act, provinces will only then make a consideration of interests. In this consideration of interests, all spatial aspects will be weighed concerning the planning area, including the aspect of the archaeology of the area, in order to reach a spatial decision. Point of departure here is that the archaeological values should in principle be preserved *in situ*. If this proves to be impossible, then excavating the area archaeologically is a good second choice (Provinciale Staten van Noord-Brabant 2002, 64; Council of Europe 1992). The aspect of archaeology is introduced through an archaeological report. In this report, the results are presented of the inventorial and valuational archaeological preliminary investigation of the planning area, on which the spatial planning decision will have to be made. The requirements to be fulfilled by these investigations are laid down in the Dutch Archaeology Quality Standard (Kwaliteitsnorm Nederlandse Archeologie, KNA[14], SIKB 2006).

4.3 The economic effect of the predictive model

By its introduction as a policy instrument by the provinces, the IKAW has greatly influenced the archaeological employment situation in the Netherlands (Rijksinspectie voor de Archeologie 2003: 9). However, the IKAW is a scientific model (Deeben *et al.* 2002), and, as is the case with all other scientific models, this model is continually subject to improvements. The archaeological boundaries will therefore inevitably shift in the future. These boundary shifts could as a consequence mean an increase or decrease in the scale of archaeological research in the Netherlands. And hence they will directly influence the number of archaeological commissions to be put on the archaeological market, which will have to be carried out by archaeological companies.

At the end of the 20th century, excavating in the Netherlands was, as a rule, a seasonal activity. The authorities mainly carried out this activity. The RACM, a number of municipalities and universities looked after the great majority of excavations in the Netherlands. At present, archaeological investigations take place all year round and are carried out for the major part by companies that depend on the supply of archaeological investigations in the excavation market. In the year 2006 for instance, there was an overstretched market situation: demand exceeds supply. The number of archaeologists supplied by Dutch universities is insufficient to meet the demand from the market. I observe that not only the mapping of archaeological boundaries of

[13] Indicative value: the probability of or predictive value for finding archaeological remains in the soil (Deeben *et al*. 2002).

[14] www.sikb.nl

unknown archaeological values has resulted in much archaeological employment in the Netherlands, but also that the drawing of these boundaries influences directly the costs to be made in the spatial planning when making and executing spatial plans. These costs will have a direct effect on the price of the structures to be developed. The desire to build in an area with a high indicative value will mean by definition that costs will have to be made for the aspect of archaeology in the preliminary investigation, and that there is a good chance – and thus a financial risk – that, prior to the execution of the building project, an archaeological excavation will have to take place. The costs of an archaeological excavation can form a substantial part of the total building costs of a project (Anonymous 2005a, 3). Besides this obligation to have an archaeological investigation undertaken, the Dutch government has also introduced the 'developer funded' principle for the Archaeological Heritage Management: the initiator pays. Both principles have led to archaeological heritage having become calculable in a building project and has thus become a financial-economic consideration in the realization of building projects. With making the archaeological heritage a monetary issue (excavation costs), the chance has increased that the objective of the preservation of the archaeological heritage *in situ* can be realized in a market economy, where in principle every piece of land has its price (Berben 2003, 4; Anonymous 2005b). And this kind of conservation is exactly what the European treaty regarding the preservation of archaeological heritage wants to accomplish. Precisely to this objective much weight is given in the Valletta Convention, because this will form a guarantee that the archaeological heritage will remain present in the ground as tangible proof, to be studied by future generations (Council of Europe 1992, Articles 2 and 4).

Due to the fact that archaeological predictive maps form the beginning of the archaeological investigation process, this means automatically that whoever controls this instrument, can also control the costs of the archaeological aspect of a building project. A shift in the boundaries of archaeological expectation areas directly influences the costs to be made within the spatial planning in the making and execution of spatial plans. These archaeological boundary shifts on the archaeological predictive maps, which are used as policy instruments, will also influence the archaeological employment situation in the Netherlands.

4.4 The archaeological inside and outside world

In the archaeological outside world, archaeological boundaries on predictive maps are seen as static for the time being. The archaeological inside world, however, has known for years that these boundaries are flexible. We can already state that more detailed predictive maps almost always lead to a larger surface of areas with a high or medium indicative value. And a larger surface area of these areas means more archaeological preliminary investigations and thus more supplementary commissions for business. The result is that the archaeological market can thus be increased.

The number of people presently employed in the Netherlands in the field of archaeology has in a short time increased nearly five-fold, to more than 1000 due to making archaeological investigation by provinces compulsory (Fokkens 2005, 3; Rijksinspectie voor de Archeologie 2003, 9). If however a trend can move one way, it can of course also move the other way. In view of the number of people now employed in the field, there is more at stake than solely the archaeological interest. The change in the Monuments and Historic Buildings Act and the new Law Spatial Planning will mean that the responsibility for the producing – and legally determining – of archaeological predictive maps will move from the provinces to the municipalities.

As soon as the archaeological outside world discovers that the archaeological boundaries are flexible, the instrument of the expectation map will inevitably be launched as a cost-saving instrument. Variables, confidence limits, chances of being hit and assumptions will change. What we are waiting for is the moment that the more detailed predictive maps – which, due to the change in the law, will be drawn up by municipalities and initiators of large projects – will lead to a larger area of archaeological expectation areas with a low archaeological expectation value. And a larger surface area of these areas will mean fewer archaeological preliminary investigations and thus fewer supplementary commissions. The result of this will be that the archaeological market will shrink.

Those who have the opportunity to make archaeological predictive models will therefore have a powerful and precious weapon in their hands. Knowledge of GIS and/or geo-statistics is actually not required for making such predictive models. At the moment anyone in the Netherlands is allowed to carry out an archaeological desktop study, so anyone is allowed to make an archaeological predictive model.

4.5 Possibilities in the field of regulatory measures: the balance of power

The question whether the makers of archaeological predictive models should meet certain requirements, and if so which ones, is perfectly valid. This question is expedient especially now as predictive maps can be powerful and at the same time 'valuable' instruments[15]. Yet this question does not in fact go far enough, in view of the implications – as discussed above – of archaeological predictive models. What of course matters in the end is the quality of the archaeological predictive models, not that of the makers, as evidently well-trained makers are required for good predictive models. But in the question as formulated at present, the makers of the expectation models are the focus and not the product. I would argue for not focusing on the maker but on the product.

I have already established above that the archaeological commercial community, and hence also the predictors employed by these companies, depend on the commissions from the market with regard to their job. In this market situation, it is in my view too dangerous to allow a caste of archaeological IT specialists to come into existence, which albeit would have to meet high standards but which would also be allowed to have a monopoly on making predictive maps.

My postulate is that archaeologists are not better people than people in other professional groups. Apart from the fact whether it would actually be possible to allow such a new 'caste of priests' to come into existence in present-day Europe, we will have to ask ourselves the question whether this is desirable. Although the endeavour is sincere, as archaeological world we should never let this work be monopolised. If only from a standpoint of development and quality contol we ought not to want this happen. All research topics that are too isolated carry the risk to come to a standstill. I would plead for the setting up of a system of 'balance of power', a system in which the controlling body and the line of business balance each other out, and keep each other sharp. A balance of power, through which quality improvement can take place on all fronts, both quality-wise and personnel-wise. At the same time, I would argue that the construction of archaeological predictive models be included as a separate issue of the Dutch Archaeology Quality Standard, and that in large spatial projects the construction of a predictive model will compulsorily be accompanied by a controlled sample[16]. This controlled sample serves as test (falsification or verification) of the formulated predictive model.

In a system of balance of power in which the one power is inevitably the archaeological commercial community, the other power will inevitably have to be, in my view, a government institution that is also able to operate independently of the market. Why independently? By definition we cannot expect the archaeological commercial community to be independent. And now that all large university archaeological institutes have an archaeological foundation that is dependent on commissions from the market, and the universities have thus become structurally dependent on this non-government funding, the universities have, as has the commercial companies, the appearance against them. That is the reason why also Dutch universities can no longer take on this independent role. Besides the archaeological commercial community, also universities have a financial advantage in making the areas in which archaeological research has to be carried out obligatory as large as possible in order to maximize the chance of supplementary commissions.

4.6 Improvement of archaeological predictive models

Would it be possible in Dutch archaeology to lay an extra controlling task with a government institution that is also capable of operating independently in the market? Apart from a possible newly-to-be-founded institution,

---

[15] This question was therefore justly raised by the organizers of the symposium *Archeologische Voorspelling en Risicobeheersing* (Archaeological Prediction and Risk Management), held at Leiden University the 1st and 2nd of March 2006.

[16] www.sikb.nl

this independent archaeological institution could in my view only be the The National Service for Archaeology, Cultural Landscape and Built Heritage (RACM – Rijksdienst voor Archeologie, Cultuurlandschap en Monumenten). Why?

On the one hand because the RACM does not really have a commercial interest in an enlargement of the archaeologically likely areas, on the other hand because the RACM – as executive department of the Ministry of Education, Culture and Science – is also in the position to determine with large projects, via the Environmental Impact Assessment (EIA in Dutch MER - Milieu Effect Rapportage) procedures, which archaeological data should be part of the to be drawn up EIA (Eerste Kamer der Staten-Generaal 2006, Article III, A to E). It is precisely with these large projects that new predictive models are commonly formulated.

In addition, with these projects in particular are the interests and opportunities greatest to deploy an archaeological expectation map as a financial regulatory instrument. The RACM could as a norm demand from EIA compulsory projects to supply a detailed predictive model including one or more controlling samples to test this model. These controlling samples should be carried out in all parts of the planning area in order to improve the (newly) drawn-up archaeological expectation model.

4.7 A plea for an independent and controlling archaeological knowledge centre

If a system could be realized within Dutch archaeology, in which a controlling government body itself would be able, and has been charged with formulating, prescribing (including a controlled sample) and verifying predictive models, this will have two big advantages for the existing archaeological order. On the one hand, it will break through the present self-fulfilling prophecy in Dutch archaeology that areas with a low indicative value will not be investigated and will become, from an archaeological point of view, emptier and emptier, and planologically more and more filled up, as it is cheaper to develop spatial plans in an area with a low indicative value. On the other hand, the know-how of making and controlling predictive maps will (also) remain in the hands of an independent body.

To make a controlling sample compulsory with large projects (EIA projects) in all areas with an archaeological indicative value has two big advantages. This compulsory sample will not only result in that the detailed predictive models becoming more underpinned and justified, but will also have the result that the development of the national Indicative Map Archaeological Values (IKAW) will not have to come to a standstill through a lack of archaeological information from areas which on the basis of the same predictive model are exempted – already sometimes for more than six years − from systematic archaeological research (areas with a low indicative value). That fieldwork for the benefit of these samples, which will be carried out in selected areas, will be carried out soundly has already been regulated in the new disposition of the Dutch Archaeological Heritage Management. The Heritage Inspection will oversee this (Rijksinspectie voor de Archeologie 2003: supplement 2)[17].

Companies will produce new archaeological predictive models for the benefit of EIA compulsory projects. However, in view of the big spatial and financial consequences of such a model, I am in favour of such a model always being recalculated and checked – in view of the large scale – by the RACM. Why would the authorities ask for an auditors' certificate for settlement of government subsidies of € 20,000 or more and not ask for a 'declaration of verification' in calculations of areas where archaeological investigation should take place. The research costs arising from this easily exceed € 20,000.

Such an independent institute could also provide second opinions for, for instance local authorities. Due to the changes accepted in the Monuments and Historic Buildings Act in 2006, in which local authorities will be forced to take into account the archaeological monuments either present in the ground or to be expected, and thus are stimulated to formulate an archaeological policy, notably local authorities will in future become large customers of predictive models. Instead of one national predictive model, a mosaic will be created in the Netherlands of local authority predictive models. Many local authorities will not have the know-how of

[17] www.Erfgoedinspectie.nl

judging these models on their merits. Both for local authorities, and for the sake of archaeology, the presence of an independent government body in the existing archaeological order, which can pass an independent judgement on the newly formulated predictive models, could be a large cornerstone for the confidence in the archaeological branch. In particular because the legislator has not only made it compulsory for local authorities to take into account the archaeological values actually present in the ground, but also the archaeological values to be expected in the ground. This latter category has however not been defined by the legislator. It is therefore unclear when an area is expected to be archaeologically valuable *c.q.* when it is a question of to be expected archaeological monuments.

In the field of the environment, for years we have known in the Netherlands such an independed institute: the National Institute for Public Health and the Environment (RIVM –Rijksinstituut voor Volksgezondheid en Milieu). This government institute not only looks after the information and the monitoring of the Public Health and Environment policy at a national level in the Netherlands, but also provides the scientific underpinning of this policy in the Netherlands. The recent change in the Monuments Act will lead to the Minister of Education, Culture and Science to be involved with the National Environmetal Policy Plan and with the RIVM report to be submitted every four years. This Minister has been given the authority to designate government bodies which should be involved by the RIVM in the drafting of the RIVM report, which in turn will play an important part in the preparation of the National Environmetal Policy Plan (Tweede Kamer der Staten-Generaal 2006: Article III, B). It is obvious that the Minister will designate the RACM as government body that will have to be brought in by the RIVM.

The RACM should develop into a knowledge institute equal to the RIVM but then for the monument, landscape and archaeology policy. This Dutch state institute would be responsible for the scientific underpinning of the monument, landscape and archaeology policy in the Netherlands and would also provide the information and monitoring of this selfsame policy. An important part thereof should be the testing of the quality of the archaeological predictive models, because these models form the basis of the execution of archaeological investigations in the Netherlands.

This would also be in line with the view of Dutch parliament that – by passing amendment 30 to the change of the Monuments and Historic Buildings Act – has pronounced on the scientific nature of archaeology. Parliament is of the opinion that archaeological investigations in relation to excavations and the execution of excavations should satisfy the requirements of scientific prudence and scientific relevance (Tweede Kamer der Staten-Generaal 2006). In the explanation of this amendment it is explicitly stated that this amendment to the law implies that scientific criteria shall be the basis of the Dutch archaeological policy. The basis of the archaeological policy is in most cases an archaeological predictive map. This deserves therefore to be launched with the utmost prudence.

References

Anonymous 2005a. Bouw vreest nieuwe Monumentenwet, Procedures langer, prijzen hoger door noodzaak van bodemonderzoek, *Het Financieele Dagblad*, 15 augustus 2005, 3

Anonymous 2005b. Stenografisch verslag van een wetgevingsoverleg van de vaste commissie voor Onderwijs, Cultuur en Wetenschap over archeologische monumentenzorg, 31 januari 2005, 14.00-16.46, *Ongecorrigeerd stenogram, Stenografisch verslag van een wetgevingsoverleg van de vaste commissie voor Onderwijs, Cultuur en Wetenschap 2004-2005, 35*

Berben, A. 2003. Goed omgaan met archeologie: een kwestie van beschaving, interview staatssecretaris Medy van der Laan over Malta-wetsvoorstel, *Malta Magazine* 4, oktober 2003, 2-4

Council of Europe, 1992. European Convention on the Protection of the Archaeological Heritage (Revised). *European Treaty Series* 143, Valletta

Deeben, J., D. Hallewas, J. Kolen, and R. Wiemer 1997. Beyond the crystal ball: predictive modelling as a tool in archaeological heritage management and occupation history. In W. Willems, H. Kars, and D. Hallewas (eds), *Archaeological Heritage Management in the Netherlands. Fifty Years State Service for Archaeological Investigations*, 76-118. Amersfoort: Rijksdienst voor het Oudheidkundig Bodemonderzoek

Deeben, J.H.C., D.P. Hallewas and Th.J. Maarlevelt 2002. Predictive modelling in archaeological heritage management of the Netherlands: the indicative map of archaeological values (2nd generation). *Berichten ROB* 45, 9-56. Amersfoort: Rijksdienst voor het Oudheidkundig Bodemonderzoek

Deeben, J.H.C. (ed.) 2008. *De Indicatieve Kaart van Archeologische Waarden, derde generatie*. Rapportage Archeologische Monumentenzorg 155. Amersfoort: RACM

Eerste Kamer der Staten Generaal, 4 april 2006. Wijziging van de Monumentenwet 1988 en enkele andere wetten ten behoeve van de archeologische monumentenzorg mede in verband met de implementatie van het Verdrag van Valletta (Wet op de archeologische monumentenzorg). *Eerste Kamer, vergaderjaar 2005-2006, 29 259, A*

Fokkens, H. 2005. *Voorbeeldige voorouders, graven naar de ideeënwereld van prehistorische boeren gemeenschappen*. Rede uitgesproken bij de aanvaarding van het ambt van hoogleraar op het gebied van de Europese Prehistorie aan de Universiteit Leiden op 15 november 2005

Leusen, P.M. van, J. Deeben, D. Hallewas, P. Zoetbrood, H. Kamermans and Ph. Verhagen 2005. A Baseline for Predictive Modelling in the Netherlands. In M. van Leusen and H. Kamermans (eds), *Predictive Modelling for Archaeological Heritage Management: A research agenda*. Nederlandse Archeologische Rapporten 29, 25-92. Amersfoort: Rijksdienst voor het Oudheidkundig Bodemonderzoek

Provinciale Staten van Noord-Brabant, februari 2002. *Brabant in Balans, Streekplan Noord-Brabant 2002*, ’s-Hertogenbosch

Rijksinspectie voor de Archeologie 2003. *RIA-jaarverslag 2002*, Zoetermeer

SIKB 2006. *Kwaliteitsnorm Nederlandse Archeologie (KNA). Ontwerp herziening versie 3.1.* Gouda: SIKB

Tweede kamer der Staten-Generaal 3 april 2006. Amendement van het lid Vergeer ter vervanging van dat gedrukt onder nr. 23 1, *Tweede kamer, vergaderjaar 2005-2006, kamerstuk 29 259, nr. 30*

5. Archaeology as a risk in spatial planning: manoeuvring between objectivity and subjectivity

René Isarin[18], Philip Verhagen[19] and Boudewijn Goudswaard[20]

5.1 Introduction

The recent revision of the Dutch Monuments and Historic Buildings Act, which implements the ratification of the Valletta Convention by the Dutch parliament has left the archaeological sector in a somewhat confused state, even though not all archaeologists recognize and accept this. There are three major changes resulting from the new legislation. Firstly, archaeology is now part of a larger democratic process of decision making, in which it is only one of many spatial factors to be taken into account. It is treated just like soil, water and air quality, ecology and noise pollution. Secondly, this transformation from an inward-facing and 'sectoral' attitude to archaeology to an integral spatial planning approach is accompanied by a shift from purely academic to more practical 'public' archaeology and a change from government-based funding to a commercial, market-based system. Furthermore, decision-making has shifted from the national and/or provincial level to the local, municipal level. The archaeological sector still has to come to terms with this new situation, and the resulting confusion is mainly felt by civic initiators and contractors who now are officially and legally obliged to deal with and take care of archaeology in their specific development area.

The primary reason for the current confusion is the fact that the Dutch national government deliberately chose not to prescribe quantitative and qualitative archaeological norms. There are no norms to decide what kind of archaeology is important, rare and worth preserving, or to what level of detail excavation data should be analysed and reported. More often than not, decisions on these issues are based solely on expert judgement, instead of on objective and predefined criteria. This subjectivity is the source of many risks in archaeological heritage management for civil initiators, as it may seriously affect the time and costs involved in dealing with archaeology.

In this chapter, we will highlight some of the risks in present-day Dutch archaeological heritage management. We stress that we will consider the risk from the viewpoint of the civil contractor or initiator of a specific spatial development, and not as the risk for the archaeological remains in that specific area. We will focus on the risks related to the phase of inventory research, and will discuss possible solutions for risk management that may be found (1) in the use of predictive modelling and (2) in the necessary development of reliable core sampling survey strategies.

5.2 The process of archaeological heritage management in the Netherlands

The process of archaeological heritage management (AHM) in the Netherlands is now generally accepted and common practice for archaeologists. It is designed to ensure that archaeology is integrated in spatial planning in an early stage. Activities potentially threatening the archaeological heritage (*i.e.* all activities likely to disturb the soil, like the construction of houses) are accompanied by archaeological research from the start. Figure 5.1 shows the various steps that have to be taken in order to arrive at a decision on what to do with archaeology. It can be seen as a process of stepwise intensification of archaeological research. Starting out with a desktop study of the complete area under development, in each subsequent step decisions are made on if and where to intensify research. This intensification moves from reconnaissance survey (most often by means of core sampling) to trenching campaigns. The latter, more detailed investigations will only be done in the areas that were decided to be archaeologically 'valuable' in the preceding step. This 'zooming in' on the areas of interest will then lead

[18] Past2Present-ArcheoLogic, Woerden, the Netherlands.

[19] ACVU-HBS, Amsterdam, the Netherlands.

[20] Past2Present-ArcheoLogic, Woerden, the Netherlands.

to a final valuation of the archaeological remains found. By using a multi-criteria decision making framework (SIKB 2006), the results of the archaeological research are evaluated. The horizontal and vertical dimensions of the site and its intrinsic value must be clear, and on the basis of the valuation a decision is made on how to preserve the valuable archaeology present: by mitigation, excavation, or supervision[21].

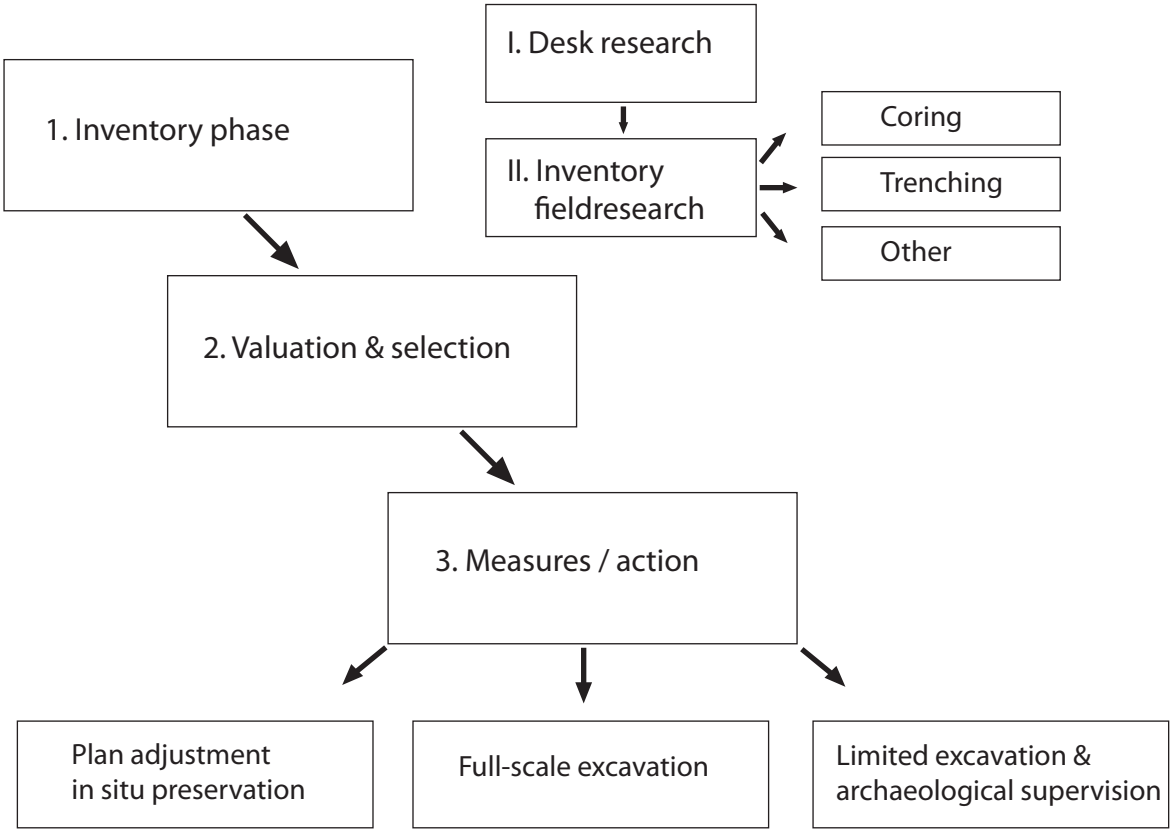

Figure 5.1 The various steps that have to be taken in order to arrive at a decision on what to do with archaeology.

We stress that this concerns *a* decision: due to the lack of norms on how to deal with archaeology (see next section) there is no single possible outcome of the decision-making process. The valuation scheme used is far from objective or transparent. Furthermore, the criteria and norms used for deciding on where to intensify research (*i.e.* in the stages before valuation) are not very well defined either. Instead, decisions are arrived at through negotiation and will inevitably result in a selection as not all the valuable archaeology can or has to be preserved *in situ* or excavated. A research agenda may serve as a policy instrument to include or exclude specific archaeological periods or research themes for the next 5 years or so, or for a specific project. But since well-defined research agendas are at the moment virtually non-existent, selection is in many cases based on the judgement of the archaeologists employed by the authorities to execute the Monuments and Historic Buildings Act.

[21] Supervision (or monitoring) is a cheap alternative to excavation, comparable to the watching briefs in English archaeological heritage management.

5.2.1 A lack of norms

Given the general absence of objective and transparent valuation criteria, it can be hard for initiators to deal with archaeology. After all, the revised Monuments and Historic Buildings Act does not prescribe any specific norms for protection of the archaeological heritage. This situation closely resembles the way in which environmental issues were incorporated in spatial planning in the late 1980s and early 1990s. Questions arose on how to measure pollution and assess its potential hazard for public health, and it was only after a decade of debate that norms were established and incorporated in daily practice. Nowadays, archaeologists have to answer similar questions about the value of archaeological sites: is it worth spending money on, and if so, is it better to excavate or preserve?

We already showed that the AHM process consists of several decision-making moments. In practice, we can distinguish four critical steps (figure 5.2), all of them potentially giving rise to debate and conflict. The first one is the decision whether archaeology is in fact present at a specific location. This decision is based on the results of desk-top study and reconnaissance survey. The second step is the decision on the size and value of the archaeological remains under consideration. This is based on the multi-criteria valuation scheme mentioned in the preceding section. Thirdly, a decision has to be made on how to realize site preservation: mitigation, excavation and/or supervision. And finally, a decision has to be made on the extent of the selected preservation measures. The level of detail of analysis and publication can be basic, but on the other hand an excavation may harvest enough data for someone to write a PhD thesis!

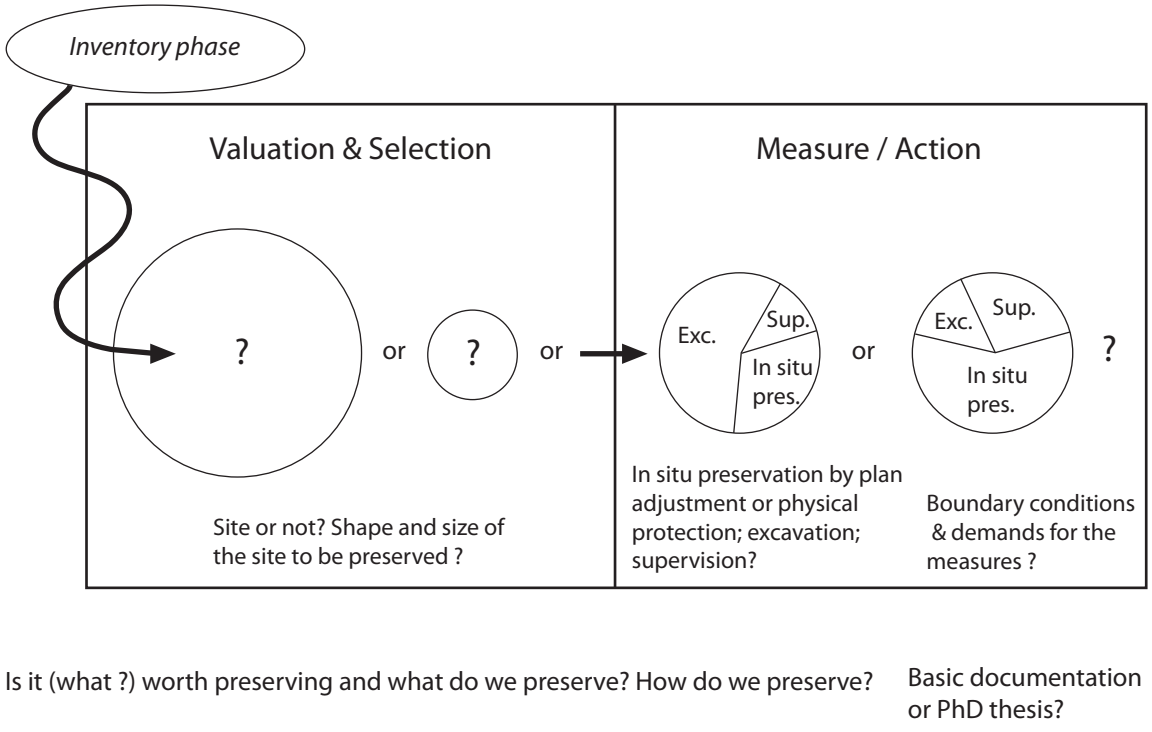

figure 5.2 The four critical steps of decision-making in the archaeological heritage management process.

In practice, the currently employed non-normative system frequently evokes debate between the 'antagonists' in archaeological heritage management: the initiators of spatial developments and the municipal, provincial or national authorities. The developers will benefit from clearly predefined and objective criteria in order to control and direct time and money in a development project. After all, archaeology is only one of the conditions they have to deal with. The authorities on the other hand have the legal obligation to protect the archaeological heritage, and will have to provide the developers, preferably beforehand, with norms for dealing with archaeology in a specific project. This clearly may lead to conflicts of interest and to discussions on the validity of the outcome of the valuation. This is exacerbated by the fact that the archaeological companies doing the research also bring their own opinions. Commercial advisors acting as mediators between developer and authorities are also adding to the debate. Obviously, the different value systems employed by the various parties involved in archaeological heritage management can easily lead to miscommunication and emotional debates.

For this reason, we feel that we should try to minimize the grounds for misunderstanding, by introducing more objective and transparent valuation criteria, and clear norms on where to draw the line between 'important' and 'not important', thus between valuable and not valuable.

5.2.2 Dealing with risk in archaeological heritage management

In the remainder of this paper, we will focus on the first decision-making step of deciding whether we are dealing with archaeology or not. It is probably the most debated and crucial issue in Dutch archaeological heritage management at the moment: decisions made in the beginning of a project cannot easily be reverted in a later stage. Furthermore, we feel it is an issue where recent scientific research in especially predictive modelling and archaeological survey has come to a stage where we can actually start to implement the desired objective decision-making criteria in practice. In order to do so however, we will first have a look at the concept of risk in archaeological heritage management.

We can look at risk from two different angles: first of all, it can mean the risk that archaeological remains are destroyed without any form of intervention. This is what most archaeologists will understand by risk in the context of archaeological heritage management. However, from the point of view of the developer, there are very different risks involved. Firstly, there is the risk of delay of the development plans. Carrying out archaeological research and obtaining permits from the authorities takes time. Secondly, there is the financial risk: if (unexpected) archaeology is present in the development area, the developer may have to pay for more research than anticipated. In practice, developers do not have many options to control these risks. They are dependent on the authorities for obtaining permits, and given the absence of norms, the authorities can pretty much do as they like, in some cases downright obstructing development plans or forcing developers to carry the costs of very expensive research. In our view, the key issue is that no one seems to be able to tell whether the (perceived) archaeological risk justifies the decisions made by the authorities.

If we look at the tools currently available to control the archaeological risk, we have to conclude that these are not very well suited for an assessment of either the archaeological or the developer's risk. Predictive models are employed to enforce survey in medium and high probability zones, but the models used do not say anything about the potential number and nature of the archaeological sites that may be found. So, the developer will only know that a survey needs to be carried out, but not what the result of the survey may be in financial and temporal terms. Similarly, core sampling is often enforced as the survey method of choice for reconnaissance survey, but without specifying the probability that certain types of sites may be missed. Furthermore, it doesn't take into account the possibility that the archaeological indicators found are not sites at all, but for example, reworked artefacts. As a consequence, survey results may show site contours that have little meaning, and in later phases sites may pop up that were not detected during survey. In those circumstances, the authorities will usually demand for new research. The key question is: who is responsible for the delay and costs?

5.2.3 Predictive modelling and risk assessment

To make predictive models a more useful tool for risk assessment, we have to stop being vague about the meaning of low, medium and high probability. A predictive map will only tell whether survey is necessary or not. It is a norm, based on (usually) an expert judgement assessment of the relative density of archaeological sites. It does not tell the developer how much of the area surveyed will be selected for further investigation. So, the only risk that can be established with some reliability is the amount of time and money that will go into reconnaissance survey.

Actually calculating the potential number and nature of sites that may be encountered during survey is far from trivial, but it is not impossible. Recent developments in statistics like resampling (see chapter 8) and Bayesian statistics (see chapter 9) now enable us to get a firmer grip on the numbers involved, and the associated uncertainties. Unfortunately, these techniques still need further study and development before they can be implemented in practice.

We can however also use a more pragmatic approach by analysing the surveys done in the past, and calculate the area that was selected for further investigation. Some of these data were collected by us, and show that in a sample of 23 projects, 23.9% of the area surveyed was selected for further archaeological research. Students from the University of Groningen (Schepers and Vosselman 2005) did a similar exercise for the province of Drenthe in the years 2003-2004. They concluded that for high probability areas 23% of the area surveyed was selected for further research; for low and medium probability areas this was 18%. While this is useful information in itself, a bandwidth for these figures would even be more helpful. For our own data, we calculated that there is a 97.5% probability that the area selected for further research will be less than 37.2% (see chapter 8, 111).

Obviously, this is only a first step towards a financial risk assessment that will be helpful to the developers and provide a necessary counterweight for the more subjective and emotional alternative. For that, we also need more information on the actual costs of dealing with different aspects of archaeology, and an assessment of the probability of particularly expensive types of research being necessary. This in turn implies that risk assessment studies must be carried out at a wider scale than the current development project, as we need to have comparative data. Who should pay for this kind of research, and how do we make the necessary data available?

5.2.4 Interpreting survey results

The principal method used for reconnaissance survey in the Netherlands is core sampling. For many years, core sampling was applied without a clear idea of its limitations. Research by Tol *et al.* (2004) shows that core sampling will never guarantee a complete detection of archaeological sites, because of its restriction to very small sampling units that are relatively widely spaced (see also Verhagen 2005). Core sampling survey results can be manipulated by changing the density and configuration of the sampling grid, and by taking smaller or larger cores. However, sites that are characterized by a low density of artefacts are typically very difficult to detect. Furthermore, core sampling survey will not always be able to tell whether the artefacts encountered are ‘in situ’. So, core sampling may both under- and overestimate the actual extent of archaeological remains in a study area.

From the point of view of the developer this is hardly satisfying, because it will cost time and money in both cases. Undetected sites that pop up in a later stage will usually prompt the authorities to demand additional research, whereas erroneously interpreted non-sites will waste precious research money and time.

The problem of non-detection is extensively discussed by Tol *et al.* (2004). They suggest that the choice for a particular survey strategy should be based on a hypothesis about the type of site that can be expected in a study region. This is called a ‘specified prediction’, and determines what survey method should be used, as not all sites are equally easy to detect. Their suggestion has currently been added as a guideline to the Dutch Archaeology Quality Standard for Archaeology (KNA version 3.1; SIKB 2006). Using this guideline, we can judge the probability of detecting site type A or B using a specific survey method. For example, if we want to

have an 80% probability of detecting a medium-sized Stone Age site, we need a 20x25 core sampling grid, using a 15 cm diameter core, and a 3 mm sieve for detection of the flints (Tol *et al.* 2006).

Despite this important step forward, a *norm* for detection probability is still missing: is 80% an acceptable limit for the specific authority (municipal, provincial, national) and developer? After all, it implies that 20% of the sites we are looking for will not be detected. And if we accept an 80% detection probability, does this mean that we will not spend any money even when we find the other 20% in a later stage of the development plan? So, while using objective and transparent criteria is necessary, establishing norms based on these criteria is even more crucial. *Most* crucial however is accepting the possibility that archaeology may be missed and, despite this, accepting the fact that the developer is not liable for the consequences.

Even then, the 80% detection limit only tells us that we will be able to detect the expected artefact concentration 4 out of 5 times. It will not allow us to correctly delimit a site. When we have struck an artefact, this will be a reason to look closer, by taking more samples in the vicinity of the find location and trying to establish a site contour in this way. This approach is also known as adaptive sampling (see Orton 2000). However, since we are dealing with imperfect detectability of the artefacts, the neighbouring samples will also be empty 1 out of 5 times – and this is assuming that we are still dealing with the same artefact density. So here we have a classical Catch 22-situation: we do not actually know what artefact density we are dealing with, so how can we be sure that a non-artefact observation is proof of the absence of a site?

In fact, the only reliable method for establishing site contours is trial trenching. Yet site contours are still drawn on the basis of core sampling surveys as if they constitute real boundaries, and trial trenching, when advised, is usually limited to those zones. Furthermore, in some cases artefact concentrations may not even be sites at all. A flexible approach should be applied instead: we should dig the trenches as far as is needed. In some cases this will be a more limited zone than the survey contour indicates, because the artefacts found were not an indication of an archaeological site. In other cases it will be more extended, because the site contains features that were not detected during survey. However, this will put the developer in a difficult position, as it means that a 'worst case' scenario will have to be adopted in order to assess the risks involved. It also complicates the situation for archaeological companies doing the research, as they will have to take into account the possibility that they will only have to do a small part of the original project proposal. And they will also have to evaluate the results of their research in the field, and keep in close contact with the developer and authorities during the fieldwork.

Bayesian statistical methods may be helpful in this context for establishing the risks involved. Nicholson *et al.* (2000) discuss the problem of trying to estimate the risk that archaeological remains of a certain size may be missed given a specific research intensity. For example, when using classical statistical methods, the probability of not detecting remains with a size of 1% of the study area (*e.g.* a 100 m^2 site in a 100x100 m survey area) is still 61% when taking 50 samples. However, since we have started our survey with a specific hypothesis in mind about the type of sites we're looking for, we might as well use Bayesian statistics to come up with a more realistic estimate. For that, we need to specify the smallest area of archaeological remains that we want to detect. The problem of imperfect detection is tackled by dividing this area by the detection probability involved. So, in the case of the medium-sized Stone Age site with a 80% detection probability, we should reduce the 'site area' to 200*0.8 = 160 m^2. We also have to specify an assessment of the probability that these remains are present at all. For the purpose of illustration, let's assume that earlier research indicated that in 10% of cases, these remains were actually found. This means that the initial probability of such sites being present is 3.7%. When taking 50 'empty' samples in the survey area, this risk is reduced to 1.0% (for the actual mathematics, see Nicholson *et al.* 2000). The risk that we missed two of these is then 0.3%. Such an approach seems helpful in analyzing the risks involved with archaeological survey, but it implies that sufficient data should be collected to estimate our prior assumptions on the presence and size of archaeological remains. The method described also has to be translated to real situations, and evaluated for its effectiveness.

5.3 Concluding remarks

No doubt, archaeology is a true risk to civil developers. In our experience, it is not the amount of money going into archaeology that most annoys the developers. Instead, they are frustrated by the fact that the 'rules of the game' are continuously changed during the AHM process, and that decisions are based on expert judgement without a clear scientific vision on the value of archaeology. As a result, the whole process may look like an endless tunnel, and archaeology is seen as a planning condition that is completely out of control. In comparison with other environmental factors like ecology, noise pollution and soil quality, archaeology lacks a clear degree of objectivity and thus professionalism. At least four non-normative steps in the AHM process can give rise to potential debate and conflicts between developers and municipal, provincial or national authorities. Decision making in archaeology is largely a subjective process.

To a certain extent, the use of expert judgement in decision making is inevitable, as not all aspects involved in valuating archaeology can, at the current state of knowledge, be translated into objective decision making schemes and norms. But even 'subjective' norms and criteria can in most cases be formulated in a transparent way. And in our view, we also have to move toward using more objective and quantitative criteria. Even at the current state of affairs, at least some objective norms can be defined at the start of a project, for example by selecting a preferred research theme based on an objective inventory of local or regional lacunae in archaeological knowledge.

Furthermore, it is essential to focus our attention on the first step in AHM of deciding whether we are dealing with archaeology or not, the most debated and crucial issue at the moment. It is not only necessary to arrive at a norm for detection probability, whether this be 70, 80 or 82.34%. We also have to learn to live with the consequences of establishing norms. This means accepting, as a rule of the game, that the developer is not liable when, because of using predefined norms, a certain portion of the archaeology is missed.

Finally, it is necessary to find financing for research that can help to control the risks involved in AHM. At the moment, hardly any funding is available for this type of research, most probably because the need for it is not generally recognized by the archaeological sector. This may to a certain extent be due to the mathematical and statistical character of this type of research – not the most sexy form of science to the conventional archaeologist. It may therefore very well be necessary to turn to the world of contractors and spatial planners to get the necessary funding.

References

Nicholson, M., J. Barry and C. 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, University College London, London
http://eprints.ucl.ac.uk/archive/00002738/01/2738.pdf

Orton, C. 2000. *Sampling in Archaeology.* Cambridge: Cambridge University Press

Schepers, M. and J. Vosselman 2005. *Archeologisch booronderzoek in Drenthe. Een onderzoek naar archeologisch prospectief booronderzoek in Drenthe in de jaren 2003-2004.* Student report, Groningen: University of Groningen
http://members.home.nl/kwassink/1.%20archeologisch%20booronderzoek%20in%20Drenthe.pdf

SIKB 2006. *Kwaliteitsnorm Nederlandse Archeologie (KNA). Ontwerp herziening versie 3.1.* Gouda: SIKB

Tol, A., Ph. Verhagen, A. Borsboom and M. Verbruggen 2004. *Prospectief boren. Een studie naar de betrouwbaarheid en toepasbaarheid van booronderzoek in de prospectiearcheologie.* RAAP-rapport 1000. Amsterdam: RAAP Archeologisch Adviesbureau

Tol, A., Ph. Verhagen, and M. Verbruggen 2006. *Leidraad inventariserend veldonderzoek. Deel: karterend booronderzoek.* Gouda: SIKB

Verhagen, Ph. 2005. Prospecting Strategies and Archaeological Predictive Modelling. In M. van Leusen and H. Kamermans (eds), *Predictive Modelling for Archaeological Heritage Management: A research agenda*. Nederlandse Archeologische Rapporten 29, 109-121. Amersfoort: Rijksdienst voor het Oudheidkundig Bodemonderzoek

6. Archaeological predictions contested: the role of the Dutch Indicative Map of Archaeological Values (IKAW) in local planning procedures

Martijn van Leusen[22]

6.1 Introduction

Predictive models may, at least in the Netherlands, play a role in the spatial planning process in the guise of maps depicting archaeological values. In such cases the predictive map is not seen as a visualisation of a scientific model, but as a planning policy map. Having been given a formal role in spatial planning procedures, the map inevitably gets drawn into conflicts between stakeholders in the future allocation of limited space – civilians, companies, and government. The attitudes of each of these segments of society towards archaeological predictions become clear when we study what happens when municipal planning policy documents get contested.

Municipal planning policies are set, and regularly updated, in formal documents called zoning schemes[23]. Before getting legal status they are tested by the next higher level of government, the provincial authorities, and if these reject – in whole or in part – a zoning scheme it often gives rise to a dispute in administrative law[24] at the Council of State[25]. During these proceedings, stakeholders can argue their objections to any part of the provincial authorities' decision that they find unjust. Inasmuch as these objections are related to the presence or absence of archaeological values, the full transcripts of cases brought before the Council of State (which can be found at www.raadvanstate.nl/uitspraken) may shed light on the actual opinions of all involved parties. This therefore presents an excellent opportunity for finding out how aggrieved private citizens and companies, as well as representatives of local and regional government, view archaeological predictive maps.

During the course of the BBO project on predictive modelling an interesting case developed in the municipality of Coevorden (Drenthe province, figure 6.1). Provincial planning authorities had refused to approve a revised zoning scheme for an area called Coevorden-Dalen, because in their opinion it provided not enough protection to zones of medium to high value depicted on the Indicative Map of Archaeological Values (IKAW) produced by the National Service for Archaeology, Cultural Landscape and Built Heritage (RACM). In the next section Ms Annet Boon (Planning Office, municipality of Coevorden) argues the case for the municipality of Coevorden. As a result of this case, which was handled on March 17th, 2006, before the Council of State, some earlier cases in which archaeological predictions were relevant to the review of zoning schemes were also looked at (figure 6.1):

1. Groningen, zoning scheme Airpark Eelde (ABRS, 22 oktober 2003)
2. Buren, zoning scheme Golf course 'de Batouwe' (ABRS, 16 juni 2003)
3. Zoning scheme Buitengebied Gemert-Bakel (ABRS, 17 juli 2002)

In section 6.3 we summarise and compare the four Council decisions. In view of the length of the full decisions and the use of specialised legal language, only the immediately relevant passages are reproduced here in Dutch (see appendix 1-4).

[22] GIA, Groningen University, the Netherlands.

[23] In Dutch "bestemmingsplannen".

[24] In Dutch "bestuursrechtelijke geschilprocedure".

[25] In Dutch "Raad van State".

6.2 The case of Coevorden-Dalen, a position statement[26]

The municipality of Coevorden is in the process of reviewing its zoning schemes, and in particular has developed a new zoning scheme for the rural area of Dalen. This new zoning scheme also makes arrangements for the protection of the archaeological heritage. The following discusses the effects of the Malta agreements on municipal zoning schemes in the Netherlands, using Dalen as a case study.

figure 6.1 Locations of the four case studies discussed in the text.

6.2.1 Archaeological maps

To work archaeology into zoning schemes, use can be made of two maps developed jointly by the RACM and the provinces: the Map of Archaeological Monuments (AMK) and the Indicative Map of Archaeological Values (IKAW) (Deeben et al. 1997; 2002; Deeben 2008). The AMK contains all listed terrains containing known archaeological remains to be protected. The IKAW attempts to depict the unknown archaeological heritage by indicating the probability that archaeological remains will be found.

6.2.2 Zoning scheme of Drenthe province

The current provincial zoning scheme for the province of Drenthe (POP II) states that the consideration of archaeological interests at an early stage in the planning process is one of the basic principles of the Malta Convention. This means that municipal planning officers, when developing zoning schemes, must take into account both known and expected archaeological values. When some spatial development is intended outside a registered building plot, and in an area containing medium to high archaeological values, a preliminary archaeological assessment (IVO) is required.

6.2.3 Archaeology and zoning schemes

The protection of archaeological values can be arranged in a zoning scheme in various ways; this depends on the type of zoning scheme, which may serve to consolidate the current situation in the municipality[27] or to prepare for new developments[28]. In the case of terrains listed on the AMK it is clear that these may be put in conservation

[26] Section 6.2 was written by Ms Annet Boon, Couvorden, planning office, 27 Februari 2006.

[27] In Dutch "consoliderend bestemmingsplan".

[28] In Dutch "ontwikkelingsgericht bestemmingsplan".

zones with associated protective measures, such as a system of construction permits[29] or a construction ban with possibility of exemption[30]. The IKAW is an uncertain factor because of its predictive character, necessitating research to establish the presence of archaeological remains that are worth protecting.

A consolidating zoning scheme is protective in character, and is not intended to enable radical new developments. In particular, there is no room for new housing developments – if these are desired, a separate scheme must be made. The preservation of archaeological remains in this kind of scheme is mainly guaranteed by the fact that no radical soil disturbing activities are foreseen. A zoning scheme aimed at enabling development, however, prepares for radical spatial developments in, for example, housing or restructuring of rural space. Here it is important that archaeological values are attended to at a very early stage.

6.2.4 Coevorden - Dalen

The figure below shows the situation of the rural zoning area Dalen, some 60 sq km in size and largely consisting of agricultural land. Small population centres and one large recreation park were excluded from the scheme.

We used the following principles in deciding on the regulations for the protection of the archaeological values in this area:

1. protection of known values (AMK terrains) and plaggen soils[31] through the exclusion of building activities outside designated building parcels, and the obligation to obtain permits for several types of soil disturbance. On plaggen soils this also serves to protect landscape values and culture-historical values.
2. in areas with a medium to high indicative value on the IKAW, spatial development is only allowed if, following preliminary archaeological assessment (IVO), an exemption is obtained or a request for modification is granted[32].
3. for the remaining areas of high indicative value, there is currently insufficient reason to insist on regulations for the protection of archaeological values. Nor is it felt that they require protection, since the zoning scheme consolidates existing land use: the current zoning poses no threat to archaeological remains that are potentially present. A further substantiation/refining of the IKAW (or, alternatively, an extension of the AMK) will be needed for this to change.

6.2.5 Conflicting views

The province of Drenthe feels that the above set of regulations does not provide sufficient protection for the archaeology in areas indicated by the IKAW. It therefore wants to impose on the municipality a system of construction permits in areas with a high indicative value[33]. The municipality, however, questions the validity of such a system for the protection of archaeological values that are not in fact *known* to be present. It is also concerned that the imposing of such a system will diminish its citizens' support for archaeological research and management activities.

The crucial point in the whole issue of withholding consent for areas with a high indicative value (other than man-made soils) is, the municipality feels, whether the presence of *indicative* values is sufficient for the imposition of a system of permits. After all, the existence of archaeological values in these areas has not yet been attested by actual research. Coevorden objects to the fact that a system of permits for any soil disturbance going deeper than 30cm would impose a heavy burden (in terms of time and costs) on the operational management of the owners and users of these areas. Areas, again, that are not *known* to contain archaeological values.

---

[29] In Dutch "aanlegvergunningstelsel".

[30] In Dutch "bouwverbod met vrijstelling".

[31] A plaggen soils is a man-made (anthropogenic) soil consisting of sods brought in from surrounding areas to improve fertility. These thick soils cover, and therefore preserve, the original surface along with its archaeological remains.

[32] According to article 11 of the Spatial Planning Act (WRO).

[33] According to article 28 of the Spatial Planning Act (WRO).

The municipality acknowledges the added value that an improved or refined map of archaeological monuments would have, but feels that this should not be linked to the *consolidating* zoning scheme Dalen – which does not allow any development - at this time. The production of such an improved or refined map would be very costly and time-consuming, and in any case would have to be made for the whole municipal territory (or better yet, the whole province) rather than for Dalen alone. The municipality believes that it is the government, and not individual citizens, that is primarily responsible for the production of such base data regarding the archaeological heritage. Only then can it expect to obtain support for the imposition of a permit system such as is already in place for terrains containing archaeological monuments.

We believe that the IKAW is not suitable for use at the municipal zoning scheme level for the following reasons:

- it is based on a limited data set, consisting of soil data and the ARCHIS database of known archaeological sites. Other relevant data, such as elevation data, cadastral minutes, aerial photographs and historical maps (also important for the precise delimitation of man-made soils) have not been used.
- the resolution of the IKAW map is too low to be used at the level of individual parcels or the scale of municipal zoning schemes.
- it remains unclear by what criteria the high, medium and low indicative values were assigned; the accompanying manual mentions no decision rules.
- it is a predictive map, that does not tell us anything about the actual presence of archaeological sites or about their size.

We therefore believe that, by withholding consent to large sections of the zoning map (including building parcels) and imposing a system of permits, the province of Drenthe is using heavy-handed and unjust means to protect archaeological values whose existence has not even been proved by actual observation.

6.3 Contesting policies based on the IKAW

6.3.1 Coevorden-Dalen

In the case of Coevorden-Dalen, a new 'consolidating' zoning scheme was made for the agricultural zone of Dalen, which allows for development within areas indicated as of high or medium value in the IKAW on condition that appropriate exemptions to or modifications of zoning decisions are obtained[34]. The province of Drenthe has checked whether this scheme complies with the Provincial Zoning Scheme (POP II) and found that, under it, IKAW zones of *high* indicative value were insufficiently protected. An additional active measure[35] was needed. The municipality feels that this is not justified in the case of *expected* (hence not *proven*) values.

Specifically with respect to the IKAW, the municipality is of the opinion that neither the mapping scale nor the construction and division of the indicative values into low-medium-high zones are sufficiently well-founded, or are simply not applicable at the scale of typical zoning schemes. These criticisms have been shown to be valid by our research group.

The municipality furthermore is of the opinion that the existence of a merely *potential* site presence, as depicted by the IKAW, is insufficient reason to protect an area with a system of permits.

The Council of State opines that the presence of areas of high indicative value requires that the municipality conduct further archaeological assessments before deciding whether special zoning regulations are needed; since the municipality has not done this, provincial authorities were justified in withholding approval. It therefore rules that a new zoning scheme must be submitted.

[34] Following procedures described in article 11 of the Spatial Planning Act (WRO).

[35] A system of permits as described in article 28 of the Spatial Planning Act (WRO).

6.3.2 Groningen airport Eelde

In the case of "Groningen Airport Eelde grounds" the province of Drenthe approved the zoning scheme submitted by the municipality of Tynaarlo, whereas appellants including IVN (Society for Environmental Education, department of Eelde-Paterswolde) allege that this cannot be justified for land parcel D2905, the ecological and culture-historical values of which would be damaged.

The Council of State agrees with the appellants' position because both a provincial report on archaeology and culture-history of plaggen soils (Spek and Ufkes 1995) and the IKAW show that, whilst this land parcel lies within an archaeologically valuable area, the lack of a preliminary archaeological assessment (IVO) means that these values are not protected by the regulations associated with the zoning scheme. Such a study should have been conducted in order to assess the significance of the potentially present archaeological values; and since it did not happen, provincial authorities should have withheld consent for this section of the zoning scheme.

In their written defence, the province of Drenthe state that it was the *relatively low value* of the plaggen soil in question that made them decide that no preliminary archaeological assessment was needed for this particular section of the zoning scheme – which was therefore approved. From this we can deduce that the province would, in the case of a man-made soil with a *high* indicative value, have withheld approval unless an assessment had taken place. In contrast, the Council opines that available data including the IKAW already show that the land parcel in question *may* be archaeologically valuable, hence an archaeological assessment exercise should have been conducted prior to the provincial review of the zoning scheme anyway.

6.3.3 Golf course 'de Batouwe'

In the case "Golfbaan de Batouwe", appellant claims that the province of Gelderland has wrongfully withheld approval to part of the zoning scheme "De Batouwe 2002" submitted by the municipality of Buren. Specifically, the province withheld approval for the part containing the *extension* of the golf course, because the municipality wrongfully did not conduct an assessment of one of two terrains of archaeological value within the extension, nor did it have the remainder of the extension assessed – for which the IKAW indicates a medium-high value. The map belonging with the current municipal rural zoning scheme "Buitengebied 1997" indicates that the area of the extension is 'of archaeological value' besides being designated 'Agricultural Zone', hence there is the additional statutory duty to conduct an Environmental Impact Assessment (EIA)[36]. The appellant argues that the municipality did, in fact, conduct sufficient research into potentially present archaeological values in the part of the planning zone involving the golf course extension.

The Council of State, however, agrees with the position of the province, namely that the medium high archaeological potential indicated by the IKAW means that the whole of the area of the extension should have been assessed, rather than just one terrain. In view of this, the province was unable to establish whether the golf course extension complies with the 1996 provincial directive on the value, protection and strengthening of archaeological values; provincial approval was therefore withheld appropriately.

This case is therefore almost identical to the previous one, except for the fact that the Council, given the existing provincial directive, now evidently considers that a *medium high* indicative value on the IKAW is sufficient to stipulate archaeological assessment as a condition for zoning scheme approval. The indicated medium high value, in this case, obviously received additional weight from the presence of two terrains of archaeological value[37].

6.3.4 Gemert-Bakel rural zoning scheme

The Council of State ruling on the zoning scheme "Buitengebied gemeente Gemert-Bakel 1998" is a rather complex one because of the large number of appellants and the diverse nature of the objections submitted. However, there was only one appellant whose objections to the approval of this zoning scheme by the province

[36] In Dutch MER - Milieu Effect Rapportage.

[37] As registered in the Central Archive of Monuments (in Dutch "CMA-terreinen").

of Noord-Brabant are relevant for our purposes. This appellant objects to the fact that his properties were co-designated an 'archaeologically valuable terrain', on the grounds that a) this designation puts too many restrictions on his agricultural practice (keeping dairy cattle), and b) the area does not in fact have an archaeological value.

The province argues that, in view of the medium high IKAW value, the finding in 1972 of Mesolithic materials, and the soil type 'plaggen soils', the co-designation as archaeologically valuable terrain with all its attendant rights and duties was justified. The Council agrees with these arguments and opines that the institution of normal exemption procedures does not unjustly burden the appellant or the requirement to obtain construction permits. It therefore rejects the appeal.

In this case, too, the medium high or high IKAW value is supported by the fact that archaeological finds had already been made in the vicinity.

6.4 Conclusion

It is important to realise, in reading about these cases brought before the Dutch Council of State, that its remit is simply to check whether provincial authorities have abided by the pertinent laws and regulations, and whether they have done so in fairness. The Council does not concern itself with the accuracy or fairness of the archaeological values expressed in the IKAW. If it is not a specific policy of a province to require further research within areas of medium high indicative value before any zoning designations may be changed, then the province cannot appeal to such a requirement when withholding all or part of a zoning scheme. Conversely, objections concerning the content of the IKAW – such as the ones the municipality of Coevorden adduces in section 6.2 above – will not be considered by the Council as long as the IKAW is a policy map by formal decree[38] of the province. It is therefore only during procedures for the determination of *new provincial policies* that an opportunity exists to successfully object to the contents and quality of the IKAW.

[38] In Dutch "vastgestelde beleidskaart".

References

Deeben, J., D. Hallewas, J. Kolen, and R. Wiemer 1997. Beyond the crystal ball: predictive modelling as a tool in archaeological heritage management and occupation history. In W. Willems, H. Kars, and D. Hallewas (eds), *Archaeological Heritage Management in the Netherlands. Fifty Years State Service for Archaeological Investigations*, 76-118. Amersfoort: Rijksdienst voor het Oudheidkundig Bodemonderzoek

Deeben, J.H.C., D.P. Hallewas and Th.J. Maarlevelt 2002. Predictive modelling in archaeological heritage management of the Netherlands: the indicative map of archaeological values (2nd generation). *Berichten ROB* 45, 9-56. Amersfoort: Rijksdienst voor het Oudheidkundig Bodemonderzoek

Deeben, J.H.C. (ed.) 2008. *De Indicatieve Kaart van Archeologische Waarden, derde generatie*. Rapportage Archeologische Monumentenzorg 155. Amersfoort: RACM

Spek, Th. and Ufkes, A. 1995. *Archeologie en cultuurhistorie van essen in de provincie Drenthe. Inventarisatie, waardering en aanbevelingen ten behoeve van het stimuleringsbeleid bodembeschermings-gebieden*. Assen: Provincial Government of Drenthe

Appendix 1 - Extract from Council of State ruling in case 200507106/1 (26 April 2006): Coevorden-Dalen

Zaak 200507106/1: appellanten tegen gedeputeerde staten van Drenthe
Datum uitspraak: 26 april 2006

Standpunt van appellanten ten aanzien van de IKAW

2.8. Appellanten stellen in beroep dat verweerder ten onrechte goedkeuring heeft onthouden aan de plandelen op de plankaart die samenvallen met de gronden die op de Indicatieve Kaart Archeologische Waarden (hierna: IKAW) de aanduiding "hoge trefkans" hebben, onder gelijktijdige oplegging van een aanlegvergunningenstelsel voor elke bodembewerking dieper dan 30 centimeter.

Appellanten hebben aangevoerd dat de in de desbetreffende gronden aanwezige waarden geen aanlegvergunningenstelsel rechtvaardigen. Een stelsel waarbij aan het college van burgemeester en wethouders een vrijstellingsbevoegdheid is verleend of een bevoegdheid tot wijziging van een bestemmingsplan met daaraan verbonden een vereiste van voorafgaand archeologisch onderzoek biedt voldoende bescherming aan eventuele archeologische waarden in die gronden aldus appellanten.

Standpunt van verweerder

2.9. Verweerder heeft de plandelen die op de plankaart rood omlijnd zijn in strijd geacht met de goede ruimtelijke ordening en heeft daaraan goedkeuring onthouden. Hij stelt zich op het standpunt dat de gemeenteraad niet heeft gemotiveerd waarom de waarden in de binnen die plandelen gelegen gronden met de aanduiding "hoge trefkans" op de IKAW niet worden beschermd tegen bodemkundige bewerkingen. Volgens verweerder dienen deze gronden eveneens te worden bestemd voor behoud en herstel van de hierin aanwezige archeologische waarden.

Voor de gebieden waaraan goedkeuring is onthouden heeft verweerder met toepassing van artikel 28, vierde lid, van de WRO, een aanlegvergunningenstelsel opgelegd. Daartoe heeft hij een voorschrift gegeven waarin is bepaald dat een aanlegvergunningenstelsel van toepassing is, zoals opgenomen in de artikelen 3 tot en met 9 van de planvoorschriften en aanvullend ten aanzien van elke bodemkundige bewerking dieper dan 30 centimeter, waartoe worden gerekend het ophogen, afgraven, woelen, mengen, diepploegen, egaliseren en ontginnen van de gronden, alsmede het vergraven, verruimen en dempen van sloten en andere watergangen en het aanleggen van drainage, voor zover het gaat om de bescherming van archeologische waarden, categorie hoge trefkans/ verwachting (buiten de essen) en zoals nader is omschreven op kaartje B1 in oranje met geruite arcering/hoge verwachting (indicatieve kaart van archeologische waarden).

De vaststelling van de feiten

2.10. Een groot deel van de gronden in het plangebied heeft op de IKAW de aanduiding "hoge trefkans".

2.10.1. Uit de plantoelichting volgt en ter zitting is bevestigd dat de gemeenteraad heeft beoogd de archeologische waarden in al deze gebieden te beschermen door alleen ruimtelijke ontwikkelingen mogelijk te maken nadat toepassing is gegeven aan een vrijstellingsbevoegdheid als bedoeld in artikel 15 van de WRO of een wijzigingsbevoegdheid als bedoeld in artikel 11 van de WRO, met de daaraan verbonden voorwaarde dat een gevraagde vrijstelling eerst kan worden verleend, dan wel het bestemmingsplan eerst kan worden gewijzigd nadat archeologisch onderzoek heeft plaatsgevonden. Het college van burgemeester en wethouders heeft in zijn

beroepschrift erkend dat verzuimd is een dergelijke regeling in de planvoorschriften op te nemen.

Uit de plantoelichting noch uit de planvoorschriften volgt dat de gemeenteraad heeft beoogd deze gebieden te beschermen tegen bodembewerkingen.

Het oordeel van de Afdeling

2.11. Het standpunt van verweerder dat de gebieden die op de IKAW de aanduiding “hoge trefkans” hebben, archeologische waarden kunnen bevatten acht de Afdeling niet onredelijk. Voorts is aannemelijk dat bouwwerkzaamheden en werkzaamheden als ophogen, afgraven, woelen, mengen, diepploegen, egaliseren en ontginnen van gronden, alsmede het vergraven, verruimen of dempen van sloten en andere watergangen en het aanleggen van drainage het bodemarchief onherstelbaar kunnen beschadigen en de daarin opgeslagen informatie verloren kunnen doen gaan.

Gelet op de stukken en het verhandelde ter zitting heeft de gemeenteraad beoogd aan de archeologische waarden in die gebieden bescherming te bieden tegen ruimtelijke ontwikkelingen (bebouwing) door dergelijke ontwikkelingen slechts mogelijk te maken na toepassing van een aan het college van burgemeester en wethouders te verlenen vrijstellingsbevoegdheid als bedoeld in artikel 15 van de WRO, dan wel een wijzigingsbevoegdheid als bedoeld in artikel 11 van de WRO.

Wat betreft bodemkundige bewerkingen bevat het plan geen voorschriften die aan de archeologische waarden in dat opzicht bescherming bieden. De gemeenteraad heeft in het besluit tot vaststelling van het bestemmingsplan niet gemotiveerd waarom geen bescherming van de archeologische waarden tegen bodemkundige bewerkingen behoeft te worden geboden, terwijl dat wel het geval is bij bouwkundige werkzaamheden. De in beroep door het college van burgemeester en wethouders aangevoerde stelling dat slechts sprake is van een verwachting van archeologische waarden en nog niet door onderzoek is aangetoond dat daarvan daadwerkelijk sprake is, is in dit verband niet deugdelijk, te minder nu de gemeenteraad zich blijkens het vorenstaande op het standpunt heeft gesteld dat aan de archeologische waarden wel bescherming toekomt. Gelet hierop heeft verweerder zich in redelijkheid op het standpunt kunnen stellen dat het besluit tot vaststelling van het bestemmingsplan in zoverre onvoldoende deugdelijk is gemotiveerd.

Verweerder heeft in de mogelijk aanwezige archeologische waarden een voldoende rechtvaardiging kunnen vinden voor het standpunt dat een aanlegvergunningenstelsel voor de bescherming daarvan tegen bodembewerkingen is aangewezen. Voor zover appellanten hebben aangevoerd dat het door verweerder opgelegde aanlegvergunningenstelsel, voor zover dit ziet op elke bodemkundige bewerkingen dieper dan 30 centimeter, een onevenredige belemmering vormt voor de agrarische bedrijfsvoering, overweegt de Afdeling dat appellanten niet aannemelijk hebben gemaakt dat dergelijke bodemkundige bewerkingen noodzakelijk zijn binnen de reguliere agrarische bedrijfsvoering.

Voor zover appellant sub 1 heeft aangevoerd dat de begrenzing van de plandelen waaraan goedkeuring is onthouden zodanig is gekozen dat die zich uitstrekt over gebieden waar geen sprake is van een hoge trefkans, overweegt de Afdeling dat verweerder in dit verband is uitgegaan van ruimtelijke eenheden waarin gronden liggen met de aanduiding “hoge trefkans” op de IKAW. Uit de bij het bestreden besluit gevoegde kaart B1 blijkt dat de gekozen begrenzing niet tot gevolg heeft dat aan op zichzelf staande ruimtelijke eenheden waarin geen sprake is van een hoge trefkans volgens de IKAW goedkeuring is onthouden, zodat geen grond bestaat voor het oordeel dat verweerder niet in redelijkheid voor deze begrenzing heeft kunnen kiezen.

Voor zover appellant sub 1 heeft aangevoerd dat de onthouding van goedkeuring aan de plandelen ten onrechte ook ziet op de bouwpercelen overweegt de Afdeling dat op de plankaart geen bouwpercelen zijn aangegeven, zodat verweerder daarmee bij de omlijning op de plankaart van de plandelen waaraan hij goedkeuring heeft onthouden geen rekening heeft kunnen houden. In de planvoorschriften zijn weliswaar voor alle bestemmingen afzonderlijke bebouwingsvoorschriften opgenomen, maar daarin zijn evenmin de locaties van de bouwpercelen aangegeven, zodat het voor verweerder in zoverre evenmin mogelijk was de gronden waarop de bouwpercelen liggen van de onthouding van goedkeuring uit te sluiten. Evenwel heeft verweerder blijkens het bestreden besluit geen bezwaar tegen de planvoorschriften inzake vaststelling van de bouwpercelen en evenmin tegen bouwwerkzaamheden op die bouwpercelen. Gelet op het bepaalde in artikel 30, eerste lid, van de WRO, dient de gemeenteraad na onthouding van goedkeuring aan het plan door verweerder, met inachtneming van het besluit van verweerder, binnen de in dat artikel genoemde termijn een nieuw plan vast te stellen. De gemeenteraad dient daarbij in acht te nemen dat verweerder geen bezwaar heeft tegen de planvoorschriften inzake vaststelling van de bouwpercelen en evenmin tegen bouwwerkzaamheden op die bouwpercelen. Wat betreft de periode dat de gemeenteraad nog geen nieuw besluit heeft genomen overweegt de Afdeling dat appellant sub 1 geen concrete bouw- en uitbreidingsplannen heeft gesteld waarvoor de onthouding van goedkeuring een belemmering vormt.

Appendix 2 - Extract from Council of State ruling in case 200205094/1 (22 October 2003): Groningen Airport Eelde

Zaak 200205094/1: appellanten tegen Gedeputeerde Staten van Drenthe
Datum uitspraak: 22 oktober 2003

Ten aanzien van de cultuurhistorische waarden overweegt de Afdeling als volgt.

Blijkens de stukken maakt het perceel D2905 deel uit van de essen ten westen van Oosterbroek. In het provinciale rapport “Archeologie en Cultuurhistorie van essen in de provincie Drenthe” uit 1995 is de archeologische waarde van dit gebied omschreven als “Es of randzone bevat een waardevol(le) archeologische vindplaats of monument”. Voorts blijkt uit de Indicatieve Kaart Archeologische Waarden van de provincie Drenthe dat dit perceel wat betreft archeologische vindplaatsen voor het grootste deel een hoge verwachtingsgraad heeft en voor een klein deel een middelmatige verwachtingsgraad.

De Afdeling stelt vast dat ten tijde van het nemen van het bestreden besluit geen verkennend archeologisch onderzoek was verricht. Weliswaar heeft verweerder de gemeenteraad verzocht dit alsnog te doen alvorens tot uitvoering van het bestemmingsplan over te gaan, maar blijkens de stukken kan bescherming van eventuele naderhand opgespoorde waarden uitsluitend plaatsvinden door overleg met de projectontwikkelaar en niet op basis van een planvoorschrift.

Mede nu aan perceel D2905 wat betreft archeologische vindplaatsen grotendeels een hoge verwachtingsgraad is toegekend, acht de Afdeling deze vorm van bescherming, zonder dat ten tijde van het bestreden besluit op basis van een verkennend onderzoek inzicht was verkregen in mogelijk te beschermen archeologische waarden, ontoereikend.

Gelet op het ontbreken van de uitkomsten van het verkennende onderzoek ontbrak bij de voorbereiding van het bestreden besluit de nodige kennis omtrent de relevante feiten en af te wegen belangen ten aanzien van de (mogelijke) archeologische waarden ter plaatse van perceel D2905.

Het beroep van [appellanten sub 5] is in zoverre en het beroep van IVN Eelde-Paterswolde en andere is geheel gegrond, zodat het bestreden besluit op dit punt wegens strijd met artikel 3:2 van de Algemene wet bestuursrecht dient te worden vernietigd.

Gelet op voorgaande behoeven de overige bezwaren van appellanten ten aanzien van perceel D2905 geen bespreking meer.

Appendix 3 - Extract from Council of State ruling in case 200206125/1 (16 July 2003): Golfbaan de Batouwe

Zaak 200206125/1: appellanten tegen Gedeputeerde Staten van Gelderland
Datum uitspraak: 16 juli 2003

2.5. Verweerder heeft reden gezien het plandeel dat betrekking heeft op de uitbreiding van de golfbaan in strijd met een goede ruimtelijke ordening te achten en heeft hieraan goedkeuring onthouden. Hij stelt zich op het standpunt dat ten onrechte geen onderzoek is gedaan naar het andere terrein van archeologische waarde binnen het deel van de uitbreiding, met cma-code 39B-A28. Verder is ten onrechte geen onderzoek gedaan naar het overige deel van de uitbreiding, waarvoor op grond van de Indicatieve Kaart Archeologische Waarden (IKAW) een middelhoge verwachtingswaarde geldt. Verweerder heeft voorts overwogen dat voor het deel van de uitbreiding met cma-code 39B-102 een m.e.r.-beoordelingsplicht geldt, nu deze gronden blijkens de plankaart van het geldende bestemmingsplan "Buitengebied 1997" zijn voorzien van de aanduiding "van archeologische waarde (AW)" naast de toegekende bestemming "Agrarisch Gebied A". Hij stelt zich tot slot op het standpunt dat in relatie tot het draagvlakcriterium een gedegen onderbouwing van de uitbreidingsbehoefte van appellante in de plantoelichting ontbreekt.

2.6. Mede gelet op het advies van de Stichting Advisering Bestuursrechtspraak voor Milieu en Ruimtelijke Ordening van 4 april 2003 stelt de Afdeling vast dat blijkens de Archeologische Monumentenkaart (AMK) binnen het gedeelte dat voor de uitbreiding van de golfbaan is bestemd één terrein van hoge archeologische waarde is gesitueerd en verder nog een ander terrein van archeologische waarde aanwezig is. Deze terreinen zijn aangeduid met cma-code 39B-102 en 39B-A28. Blijkens de stukken heeft RAAP Archeologisch Adviesbureau in opdracht van de gemeente Buren een Aanvullende Archeologische Inventarisatie d.d. 16 maart 2000 uitgevoerd, waarbij alleen het terrein met cma-code 39B-102 is onderzocht. Appellante betoogt hierbij dat verweerder in het kader van het overleg ingevolge artikel 10 van het Besluit op de ruimtelijke ordening 1985 heeft geadviseerd alleen het terrein met cma-code 39B-102 te laten onderzoeken. Verweerder stelt zich daarentegen op het standpunt dat hij tijdens dit overleg voldoende duidelijk heeft gemaakt dat naar de gehele uitbreiding van de golfbaan archeologisch onderzoek dient te worden verricht.

De Afdeling is ongeacht deze communicatie tussen appellante en verweerder van oordeel, dat verweerder zich terecht op het standpunt heeft gesteld dat het uitgevoerde archeologisch onderzoek te beperkt is, nu niet het gehele uitbreidingsgebied is onderzocht. Dit gezien de middelhoge verwachting van archeologische vindplaatsen en het andere terrein van archeologische waarde met cma-code 39B-A28 binnen het gebied. Verweerder is derhalve niet in staat geweest om te beoordelen of de gehele uitbreiding van de golfbaan in overeenstemming is met de in het Streekplan Gelderland 1996 opgenomen essentiële beleidsuitspraak voor het landelijk gebied, dat archeologische waarden van groot belang zijn en zullen worden ontzien en waar mogelijk versterkt.

Als aanzienlijke potentiële milieugevolgen, die uit milieueffectrapporten naar voren zijn gekomen, vermeldt pagina 73 van de Nota van Toelichting effecten op bodem en water, geluidhinder, verkeersaantrekkende werking, gevolgen voor het landschap, de cultuurhistorie en de archeologie alsmede gevolgen voor de flora, fauna en ecologie.

Appendix 4 - Extract from Council of State ruling in case 199903036/1 (17 July 2002): Buitengebied Gemert-Bakel

Zaak 199903036/1: appellanten tegen gedeputeerde staten van Noord-Brabant
Datum uitspraak: 17 juli 2002

2.13. [appellant sub 9] heeft aangevoerd dat verweerders ten onrechte goedkeuring hebben verleend aan het plan, in zoverre aan zijn gronden aan de [locatie sub 9] de medebestemming "Archeologisch waardevol terrein" is toegekend. Hij stelt dat deze medebestemming hem teveel beperkt in het agrarische gebruik van zijn gronden (melkrundveehouderij). Hij bestrijdt de archeologische waarde van het gebied.

Ook vreest hij schade aan zijn gewassen als gevolg van vernatting van zijn gronden, indien de slootpeilen in het gebied worden verhoogd met het oog op natuurontwikkeling.

2.13.1. Verweerders hebben geen aanleiding gezien deze plandelen in strijd te achten met een goede ruimtelijke ordening. Zij hebben in aanmerking genomen dat het gebied op de Indicatieve Kaart Archeologische Waarden met een middelhoge verwachtingswaarde is aangeduid en dat in 1972 nabij de woning van appellant vondsten uit het mesolithicum zijn gedaan. Gezien de vondsten uit het verleden, de bodemsoort (enkeerd- of esdekgronden) en de verwachte archeologische waarde, achten verweerders de medebestemming "Archeologisch waardevol terrein" aanvaardbaar.

Ten aanzien van de vrees voor vernatting, wijzen verweerders erop dat het aanleggen van drainage een oplossing kan bieden.

2.13.2. De gronden op de plankaart aangewezen voor "Archeologisch waardevol gebied", zijn daarmee (mede) bestemd voor de bescherming van in deze gronden aanwezige archeologische sporen.

Voor zover deze medebestemming een deel van het bouwblok van appellant betreft, geldt ingevolge artikel 21, lid B, een verbod deze gronden te bebouwen. Burgemeester en wethouders zijn bevoegd vrijstelling van dit verbod te verlenen, indien het archeologisch belang zich niet hiertegen verzet en vooraf een deskundig advies is ingewonnen bij de heemkundekringen in de gemeente en de Rijksdienst voor Oudheidkundig Bodemonderzoek.

Voor zover de medebestemming ziet op een 20 tot 80 meter brede strook achter het bouwblok, geldt, dat voor een aantal, in lid C genoemde werkzaamheden, waaronder het diepwoelen of diepploegen van de bodem en het aanleggen van drainage – dieper dan 0.40 m – , een aanlegvergunning is vereist.

2.13.2.1. Het standpunt van verweerders dat het gebied te beschermen archeologische waarden kan bevatten, acht de Afdeling niet onredelijk gelet op de verwachtingswaarde die de Rijksdienst voor Oudheidkundig Bodemonderzoek van het gebied heeft en de in het verleden ter plaatse gedane vondsten. Dat deze waarden door de ophoging en verharding van een gedeelte van het bouwblok (in 1976) en door de ruilverkaveling geheel zouden zijn verdwenen, acht de Afdeling, mede gelet op het deskundigenbericht, niet aangetoond. Voorts is aannemelijk dat bouwwerkzaamheden en werkzaamheden als diepwoelen en diepploegen en het aanleggen van drainage het bodemarchief onherstelbaar kunnen beschadigen en de daarin opgeslagen informatie verloren kunnen doen gaan. Het opnemen in het plan van een bouwverbod gekoppeld aan een vrijstellingsprocedure dan wel een aanlegvergunningvereiste geeft het gemeentebestuur de mogelijkheid na te gaan in hoeverre archeologische waarden door die werkzaamheden worden bedreigd en biedt het de mogelijkheid zonodig voorwaarden stellen. Dat het volgen van deze procedure een onevenredige belemmering voor de agrarische bedrijfsvoering vormt,

is de Afdeling niet gebleken. Het gebied met de medebestemming betreft slechts een klein deel van de gronden van appellant. Voorts kan het met een aanvraag gemoeide tijdsbeslag blijkens het verhandelde ter zitting beperkt blijven tot enkele dagen. Bovendien heeft het gemeentebestuur zich bereid verklaard met appellant een convenant te sluiten over onderhoud en beheer, waardoor in de praktijk geen aanlegvergunning noodzakelijk behoeft te zijn.

2.13.3. Voor zover appellant vreest voor schade aan zijn gewassen als gevolg van vernatting, indien de slootpeilen in het gebied worden verhoogd met het oog op natuurontwikkeling, blijkt uit de stukken, waaronder het deskundigenbericht, dat veel agrariërs in het plangebied in toenemende mate kampen met verdroging en dat daarom het gemeentebestuur het voornemen heeft het water langer in het plangebied vast te houden. Uitvoering van dit voornemen betekent dat in droge gebieden sloten zullen verdwijnen of zullen worden verkleind/versmald en dat in de lager gelegen gebieden – waarin het perceel van appellant ligt – sloten zullen worden aangelegd of verruimd of opnieuw gedraineerd.

Naar hiervoor onder 2.5.3. is overwogen, is voor dergelijke werkzaamheden een aanlegvergunning vereist. Appellant kan zonodig een aanlegvergunning aanvragen voor het opnieuw draineren van zijn gronden dan wel gebruik maken van de rechtsbeschermingsmogelijkheden die de wet biedt, indien hij van mening is dat afgifte van een aanlegvergunning elders in het gebied vernatting van zijn perceel tot gevolg heeft. Daarmee zijn de belangen van appellant naar het oordeel van de Afdeling voldoende gewaarborgd.

2.13.4. Gezien het vorenstaande ziet de Afdeling geen aanleiding voor het oordeel dat verweerders zich niet in redelijkheid op het standpunt hebben kunnen stellen dat het plan op deze onderdelen niet in strijd is met een goede ruimtelijke ordening. In hetgeen appellant heeft aangevoerd ziet de Afdeling evenmin aanleiding voor het oordeel dat het bestreden besluit op dit onderdeel anderszins is voorbereid of genomen in strijd met het recht. Hieruit volgt dat verweerders in zoverre terecht goedkeuring hebben verleend aan het plan.

Het beroep van [appellant sub 9] is ongegrond.

7. Testing archaeological predictive models: a rough guide[39]

Philip Verhagen[40]

7.1 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 predictive modelling for archaeological heritage management purposes, there is a fundamental problem with the currently used predictive models and maps: 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. Because of 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 and Sebastian (1988). Because of this, 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) 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 can be found in the next chapter and was recently published by Verhagen (2007).

7.2 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 also Wheatley 2003 and 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.
- 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. Archaeologists generally recognize the necessity of a satisfactory explanatory framework for a predictive model, 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.

---

[39] This chapter has been published before as Verhagen 2008.

[40] ACVU-HBS, Amsterdam, the Netherlands.

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 (see chapter 8 figure 8.1).

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[41] is often used to measure model performance, as it combines the two criteria of accuracy and precision in one, easily calculated measure. But 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. So 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: see chapter 8 figure 8.2) 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 a play a major role either. The question therefore is: can we actually establish these risks?

7.3 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 obliged, 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 have been used as well, like gain development graphs (Deeben *et al*. 1997; Verhagen and 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 not only control the risks involved to some extent, but it can also give us a baseline to compare between different predictive maps.

7.4 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 that say 70% of the known archaeological sites is 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 our 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 do not 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

[41] Specified as 1 – pa/ps, where pa=the proportion of area (precision) and ps = the proportion of sites (accuracy) covered by the tested probability zone of a predictive model (Kvamme 1988).

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 and 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 any good 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 were collected according to the principles of probabilistic sampling, we can consider the split sample to be an independent test data set. And 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[42]. 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[43]. This method was already used by Rose and Altschul (1988) and 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 and 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. Especially small sample size, small population size and the assumption of random sampling are limiting 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. It can therefore also be of interest for the development of site density estimates and associated confidence intervals as well, provided we have control over the surveyed areas.

---

[42] 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.

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

7.5 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; and
- survey methods should be chosen such that bias in site recording is avoided.

An important obstacle to this is the difficulties of collecting data from many small survey projects, such as are 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 and Churchill (2000) for an example from practice; Wheatley (2003) for a critique of this approach; and Verhagen (2005) for some cases of institutionalised 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. However, 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 do not know the proportion of the area sampled, which makes it impossible to specify statistical estimates and corresponding confidence limits of site density. And on top of that, we cannot predict the size of the area that should be sampled in order to obtain the required sample size, as long as we do not 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. We could evidently 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 sub-surface 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; and
- 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. Correcting site density estimates for imperfect detectability is relatively easy – at least in theory. 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. 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; 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, a way out 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 these 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 are 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.

7.6 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 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, M.J. 2003. *Statistics in Archaeology*. London: Hodder Arnold

Deeben, J., D. Hallewas, J. Kolen, and R. Wiemer 1997. Beyond the crystal ball: predictive modelling as a tool in archaeological heritage management and occupation history. In W. Willems, H. Kars, and D. Hallewas (eds), *Archaeological Heritage Management in the Netherlands. Fifty Years State Service for Archaeological Investigations*, 76-118. Amersfoort: Rijksdienst voor het Oudheidkundig Bodemonderzoek

Deeben, J.H.C., D.P. Hallewas and Th.J. Maarlevelt 2002. Predictive modelling in archaeological heritage management of the Netherlands: the indicative map of archaeological values (2nd generation). *Berichten ROB* 45, 9-56. Amersfoort: Rijksdienst voor het Oudheidkundig Bodemonderzoek

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

Efron, B. and R.J. Tibshirani 1993. *An Introduction to the Bootstrap*. Monographs on Statistics and Applied Probability 57. New York: Chapman & Hall

Gibbon, G.E. 2002. *A Predictive Model of Precontact Archaeological Site Location for the State of Minnesota*. Appendix A: Archaeological Predictive Modelling: An Overview. Saint Paul: Minnesota Department of Transportation
http://www.mnmodel.dot.state.mn.us/chapters/app_a.htm.

Griffin, D. and T.E Churchill 2000. *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), 137-153

Hand, D.J. 1997. *Construction and Assessment of Classification Rules*. Chichester: John Wiley & Sons

Judge, W.J. and L. Sebastian (eds) 1988. *Quantifying the Present and Predicting the Past: Theory, Method, and Application of Archaeological Predictive Modelling.* Denver: U.S. Department of the Interior, Bureau of Land Management Service Center

Kvamme, K.L. 1988. Development and Testing of Quantitative Models. In W.J. Judge and L. Sebastian (eds), *Quantifying the Present and Predicting the Past: Theory, Method, and Application of Archaeological Predictive Modelling,* 325-428. Denver: U.S. Department of the Interior, Bureau of Land Management Service Center

Kvamme, K.L. 1990. 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, Vol. 3, 257-295. Bonn: Holos-Verlag

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

Lunneborg, C.E. 2000. *Data Analysis by Resampling: Concepts and Applications*. Pacific Grove: Duxbury Press

Orton, C. 2000. *Sampling in Archaeology*. Cambridge Manuals in Archaeology. Cambridge: Cambridge University Press

Rose, M.R. and J.H. Altschul, 1988. An Overview of Statistical Method and Theory for Quantitative Model Building In W.J. Judge and L. Sebastian (eds), *Quantifying the Present and Predicting the Past: Theory, Method, and Application of Archaeological Predictive Modelling,* 173-256. Denver: U.S. Department of the Interior, Bureau of Land Management Service Center

Shennan, S. 1997. *Quantifying Archaeology*. 2nd Edition. Edinburgh: Edinburgh University Press

Simon, J.L. 1997. *Resampling: The New Statistics*. 2nd Edition
http://www.resample.com/content/text/index.shtml

Tol, A., Ph. Verhagen, A. Borsboom and M. Verbruggen 2004. *Prospectief boren. Een studie naar de betrouwbaarheid en toepasbaarheid van booronderzoek in de prospectiearcheologie*. RAAP-rapport 1000. Amsterdam: RAAP Archeologisch Adviesbureau

Verhagen, Ph. 2005. Prospecting Strategies and Archaeological Predictive Modelling. In M. van Leusen and H. Kamermans (eds), *Predictive Modelling for Archaeological Heritage Management: A research agenda*. Nederlandse Archeologische Rapporten 29, 109-121. Amersfoort: Rijksdienst voor het Oudheidkundig Bodemonderzoek

Verhagen, Ph. 2007. Predictive models put to the test In Ph. Verhagen, *Case Studies in Archaeological Predictive Modelling*. ASLU 14, 115-168. Leiden University Press

Verhagen, Ph. 2008. Testing archaeological predictive models: a rough guide. In A. Posluschny, K. Lambers and I. Herzog (eds), *Layers of Perception. Proceedings of the 35th International Conference on Computer Applications and Quantitative Methods in Archaeology (CAA), Berlin, Germany, April 2–6, 2007*. Kolloquien zur Vor- und Frühgeschichte, Vol. 10, 285-291. Bonn: Dr. Rudolf Habelt GmbH

Verhagen, Ph. and J.-F. Berger 2001. The hidden reserve: predictive modelling of buried archaeological sites in the Tricastin-Valdaine region (Middle Rhône Valley, France). In Z. Stančič and 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, 219-232. Oxford: Archaeopress

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

Whitley, T.G. 2004. Causality and Cross-purposes in Archaeological Predictive Modeling. In A. Fischer Ausserer, W. Börner, M. Goriany and L. Karlhuber-Vöckl (eds), *[Enter the past]: the E-way into the four dimensions of cultural heritage: CAA 2003: Computer Applications and Quantitative Methods in Archaeology: Proceedings of the 31th Conference*, Vienna, Austria, April 2003. BAR S1227, 236-239 and CD-ROM. Oxford: Archaeopress

8. Predictive models put to the test[44]

Philip Verhagen[45]

8.1 Introduction

8.1.1 Background

In 2002, the research project 'Strategic research into, and development of best practice for, predictive modelling on behalf of Dutch cultural resource management' (Kamermans *et al.* 2005; van Leusen *et al.* 2005) started out by identifying the research themes that were considered to be of great importance for the improvement of the quality of predictive modelling in the Netherlands. One of these themes concerned testing of currently used predictive models, as a means to assess their quality. Very little seemed to be known about the best way to test a predictive model, and in practice tests that have been carried out were limited, and have not been used in a systematic manner to improve the predictive models. At the same time, more and more data sets have become available for predictive model testing because of the enormous increase of archaeological research carried out in the Netherlands, following the ratification of the Valletta Convention. It was therefore decided that the subject should be studied in more detail in the second phase of the project. The current chapter is the result of this more detailed investigation, which has been carried out between January and July 2005.

The research questions defined for this study are:

- can we identify testing methods that can measure predictive model quality in an unambiguous way, and that will allow us to say whether model A is doing better than model B?
- if these methods exist, what kind of predictive models do we need in order to apply them?
- is testing really necessary? Can we perhaps deal with the issue of predictive model quality by defining statistical probabilities and confidence limits?
- and finally: do we have the data sets that will allow us to carry out predictive model testing in a rigorous way? And if not, how can we generate these in the future?

These questions are addressed through a review of existing testing methods for archaeological predictive models, that have appeared in and outside archaeological literature since the late 1980s (sections 8.2, 8.3 and 8.4). This analysis is followed by an exploration of the potential of currently available archaeological data sets in the Netherlands for predictive model testing purposes (section 8.5). In section 8.6, the question of suitable models will be addressed: what testing methods are applicable to different types of models, and can we identify the model types best suited for testing? In section 8.7, the conclusions and recommendations of the study will be presented.

8.1.2 A note on terminology

A test is a procedure for critical evaluation. It is a means of determining the presence, quality, or truth of something. As such, it is a central concept in statistics, where formal tests are used to compare between two samples, or between a sample and a statistical model. The goal of statistical testing is to decide whether there is a significant difference between the two, to accept or reject the 'null hypothesis' of no difference. This traditional way of statistical testing is not applicable to most predictive models. In general, predictive models are not cast into the form of a statistical model with estimates of the parameter of interest (*e.g.* site density) and the associated confidence limits of this estimate (see also section 8.4).

---

[44] This chapter has been published before in Verhagen, Ph. 2007.

[45] ACVU-HBS, Amsterdam, the Netherlands.

Instead, predictive models are usually the result of a classification procedure. Quantitative testing methods for classifications are based on the concept of correct prediction, and the term validation is often used for the comparison of a classification to the test data. In order to make this comparison, performance measures are calculated, most of which try to capture the error rate of the prediction.

In predictive modelling literature (and even in statistical handbooks), these differences are not spelled out. Instead, the terms performance, validation and testing are used in the context of predictive modelling without a clear definition of their meaning. It is therefore useful to introduce some basic definitions that will be used throughout this chapter.

- predictive model *performance* is the degree to which a model correctly predicts the presence or absence of archaeological remains. This does not mean the presence or absence of *new*, independently collected data. In fact, in most cases performance is only measured using the data set that was collected for setting up the model.
- predictive model *validation* is the act of comparing a test data set and a model, in order to establish the model's performance. Again, this does not necessarily imply the use of new data.
- predictive model *testing* is the act of comparing a test data set and a model, in order to either accept or reject the model. This can only be done using independently collected data.

The fact that predictive model performance is most of the times calculated with the data set used for setting up the model is criticized by Wheatley (2003), who states that performance must mean the extent to which a model predicts new, independently collected data. While understanding his point, and agreeing with it, this is not current practice, and in describing the methods and techniques used, I will speak of performance regardless of the data set that is used for validation.

8.1.3 Expert judgement testing: an example from practice

In order to put the issue of predictive model testing into perspective, it is useful to start with an example from current practice. A watching brief[46] performed by Lange *et al.* (2000) along the proposed gas pipeline between Andijk-West and Wervershoof (West-Friesland, province of Noord-Holland) perfectly illustrates how the process of archaeological heritage management functions in the Netherlands, and what conclusions were drawn from an 'intuitive' predictive model test. In fact, the watching brief report was recommended to me as an example of where such a test had proved the model to be wrong[47].

The predictive model used to decide what part of the pipeline was to be monitored, was based on the theory that, in this particular area, settlement (dating from the Middle and Late Bronze Age) is confined to fossil tidal creek beds, which constitute the highest parts in the landscape[48] (Ente 1963; IJzereef and van Regteren Altena 1991; Buurman 1996). Agricultural activities were located on the flanks of these creek beds. The area was restructured between 1972 and 1979 during a land redistribution program; this included land improvement by means of levelling. In consequence, it was supposed that the topsoil had been removed in most of the area, and that any archaeological remains in the restructured zone had been either severely damaged or lost.

The gas pipeline, with a total length of 7 km, had first been prospected by means of core sampling (de Jager 1999). As the core sampling survey did not reveal any archaeological finds, no prior evidence existed for the presence or location of archaeological sites on the pipeline. However, several sites were known to exist close the pipeline from archaeological research carried out during the restructuring program.

---

[46] The term watching brief is used here in the sense it is used in British archaeological heritage management: an archaeological monitoring operation during construction activities; while the term 'monitoring' is also sometimes used in this context (*i.e.* in Irish archaeological heritage management), in the Netherlands monitoring implies making observations at regular intervals on a site that is not under direct threat.

[47] Heleen van Londen (AAC Projectenbureau, University of Amsterdam), personal communication.

[48] The fossil creek beds have become the highest part in the landscape because of 'relief inversion'; while the sandy creek bed deposits have retained their original volume, the surrounding areas with clay and peat have gradually subsided because of dehydration and oxidation.

On the basis of the prospection results it was concluded that only 2.5 km of the pipeline was located in an area of high archaeological potential, and consequently the provincial authorities of Noord-Holland decided to impose a watching brief operation on this stretch only.

During the watching brief operation a substantial number of archaeological features was discovered, revealing a total of 6 Bronze Age and 2 Medieval 'habitation clusters'. The diameter of the clusters ranged from approximately 25 to 80 meters. Between these 'sites', ample evidence for agricultural activity was found in the form of ditches, plough marks and parcel boundaries. The fact that the core sampling survey did not reveal any archaeological evidence in the watched zone is not surprising. Most of the evidence found during the watching brief operation consisted of archaeological features, which are very difficult to interpret in core samples. Furthermore, the density of artefacts in Bronze Age sites in this area is usually very low (Tol *et al.* 2004)[49], even though Lange *et al.* (2000) remark that in one place a 'relatively large amount' of pottery shards was found. In total however, they described only 61 pieces of pottery from the 8 discovered sites. In addition, 199 pieces of bone and 22 fragments of stone were described. The total number of artefacts may have been higher, as it not usual practice to count artefacts that cannot be adequately described. Even when taking this into account, it is hardly surprising that the core samples did not reveal any archaeological evidence; these small amounts have very low detection probabilities when taking core samples.

Lange *et al.* (2000) concluded that the lower lying areas (the flanks of the creek beds) had to a large degree escaped destruction, and even in the higher areas a substantial amount of archaeological evidence could still be found. As the degree of destruction of archaeological features in the inspected area was much less than expected, Lange *et al.* (2000) decided to do a field survey (a 'true' watching brief) in the remaining 4.5 km of the pipeline. This resulted in the discovery of evidence for habitation in the low probability zone as well, consisting of a substantial amount of features and some finds.

The predictive model that was used to decide on the delimitation of the watching brief operation was therefore not adequate:

'The results of the core sampling survey and the associated low expectation for archaeological values in the area are not confirmed by reality' (translated from Lange *et al.* 2000, 45).

It also seems that a layer of dark soil that in places covered the archaeological features had not been interpreted during the core sampling survey as an important indication for the possible presence of archaeological remains. However, the assumption that habitation was confined to the creek beds, and agricultural activity to the flanks of the creek beds, was confirmed by the results of the watching brief operation. So, the predictive model had not been found wanting in its basic assumptions of site distribution, but in its assessment of the degree of disturbance of the area, and in its neglect of the off site zones.

Now, the interesting thing is that the prospection report by de Jager (1999) did not present a predictive map, although it does depict the extent of the sandy creek bed deposits as established by Ente (1963) as zones of potential archaeological interest. The report also indicates that archaeological features might still be preserved in the restructured areas, citing evidence from earlier prospection projects. Furthermore, it delimited several previously unknown zones with sandy creek bed deposits, some of which were located *outside* the area eventually selected for the watching brief operation. In the conclusions of the report however, it was stated:

'Two settlements known from literature are intersected (...) In between, a lower lying area is found where no traces of Bronze Age occupation are expected' (translated from de Jager 1999, 16).

The report also contained a recommendation to perform watching brief operations on almost all locations where sandy creek bed deposits had been found. The final selection of the area for the watching brief nevertheless

[49] This is attributed to the poor quality of the pottery, which will weather easily.

resulted in a more limited area. It therefore seems that the decision made was mainly based on de Jager's conclusion that no traces of Bronze Age occupation would be found outside the selected zone, and that the uncertainties with respect to the remaining zone were not seriously considered.

Obviously, the whole watching brief operation was never intended to evaluate the results of the core sampling survey, and the conclusions drawn on the survey's quality seem a bit too harsh. On the other hand, the prospection report failed to clearly indicate the limitations of the survey method chosen, and concluded too rapidly that the unwatched zone was of no interest at all.

From the watching brief report, three important conclusions can be drawn: firstly, watching briefs are a good method to evaluate predictive maps, as they uncover an uninterrupted stretch of soil, both in high and low probability areas, and permit the unobstructed observation of features and finds. A core sampling survey is much less suitable, and Lange *et al.* (2000) even recommend the watching brief as an alternative to survey, which in this case, where archaeological sites are extremely difficult to detect by other means, is a defendable position. Secondly, it can be concluded that the (political) decision on where to carry out the watching brief, was taken on the basis of de Jager's final conclusions, and failed to take into account the uncertainties in the unwatched zones. And thirdly, it is clear that the results from the watching brief cannot be taken as proof that the model was wrong, even though sites were found where the archaeologists had not really been expecting them.

The point is, of course, that a 'wrong prediction' is often seen as a prediction in which the model failed to indicate the presence of archaeology. In this way of thinking, any observation that falls outside the high potential zone is proof that the model was wrong. A test of some kind has been performed, but one that takes a very strict view of model quality: the model should be able to predict all the important archaeological remains, or it fails. In practice, no predictive model will ever be able to conform to this standard. It is therefore of primary importance that we are able to define the quality of the model, both before and after testing. For this, it is inevitable that we use quantitative methods, and that we use test sets that are sufficiently large. The Andijk-Wervershoof watching brief clearly fails on both accounts.

8.2 Model performance assessment

In this section, a number of methods will be discussed that deal with obtaining a measure of predictive model performance. Sections 8.2.1 to 8.2.4 and 8.2.6 discuss measures and techniques that have been used for model performance assessment in predictive modelling, with varying amounts of success. Sections 8.2.5 and 8.2.7 use some of these methods to judge the performance of some Dutch predictive models. In section 8.2.8 the effects of spatial autocorrelation and spatial association on model performance will be shortly investigated. In section 8.2.9 finally, the utility of the reviewed techniques for model quality assessment is discussed.

For a better understanding of what follows, a distinction must first be made between the North American and Dutch practice of predictive model construction. In the United States, predictive models are often made using samples of land parcels (or 'quadrats') that either yield a site or a non-site observation. These can easily be transferred to grid cells in a raster GIS. The modelling, often by means of logistic regression techniques, then results in two models: a site probability model, and a non-site probability model. These are then compared to decide where to place the boundary between the 'site-likely' and the 'site-unlikely' zone (see also section 8.2.4). Dutch models predict the relative density of sites in (usually) three zones of high, medium and low probability, based on point observations of sites only. This difference has consequences for the ways in which model performance can be established.

In predictive modelling literature the terms accuracy and precision[50] are often used to describe model performance. In everyday language, accuracy is equivalent to correctness or exactness. Precision can be used as a synonym for exactness as well, but it also refers to the degree of refinement with which an operation is performed or a measurement stated. In predictive modelling however, accuracy takes on the meaning of correct prediction: does the model capture most of the sites? Precision in predictive modelling refers to the ability of the

[50] Referred to as 'specificity' by van Leusen *et al.* 2005, 33.

model to limit the area of high probability as narrowly as possible (figure 8.1). Accuracy and precision together determine the performance of the model (see also Whitley 2005b): a good model should be both accurate and precise. The term 'model' in fact only refers to the site-likely or high probability zone of a two-zone model. For a three- (or more) zone model, accuracy and precision can be determined per zone, as an indication of the performance of each individual zone.

Note that the definition of accuracy and precision in predictive modelling is different from its statistical meaning. Orton (2000a) states that statistical accuracy is 'the difference between the sample's estimate and the true value of the parameter', and precision is 'a range of possible values, a confidence interval, rather than a single value'.

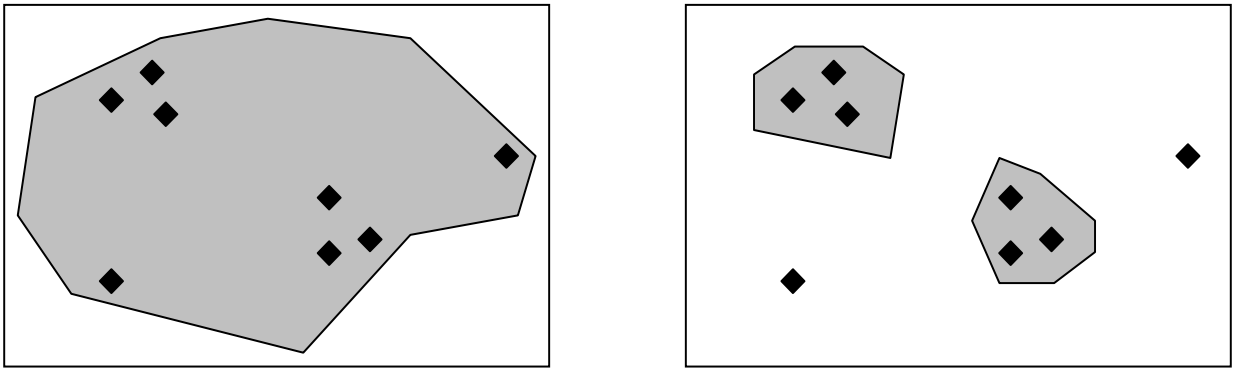

Figure 8.1 The difference between accuracy and precision. The model to the left is 100% accurate: it captures all sites (the black lozenges) in the model (depicted in grey). The model to the right is less accurate, but more precise.

Closely related to accuracy is the concept of gross error (Altschul 1988) of the model. This is the proportion of sites found in the site-unlikely zone of a two-zone model. Gross error in a model may lead to either unnoticed destruction of archaeological sites, or it can create unforeseen extra costs for mitigation measures or excavation in a development plan, as these zones will usually be without any form of legal protection. Altschul (1988) also defined the wasteful error of the model as the proportion of non-sites found in the site-likely zone of a two-zone model. A wasteful error is a less serious problem from the point of view of archaeologists, but a model that contains a large amount of wasteful error is over-protective and may lead to higher costs for developers, as larger areas will have an obligation to do survey. The concept of wasteful error is only relevant to models based on site and non-site observations. In Dutch practice, non-site observations are not used for model construction, and wasteful error can therefore not be calculated. Gross error on the other hand can be used as a measure of model quality for Dutch models, as it only implies calculating the proportion of site observations outside each probability zone.

The risk of making a gross error is inversely related to that of making a wasteful error. High accuracy in a predictive model minimizes gross error, and high precision therefore minimizes wasteful error. Furthermore, the linguistic, 'normal' definition of accuracy implies that both types of error should be minimized. Statistical literature dealing with classification error therefore combines gross and wasteful error into one measure known as the error rate (Hand, 1997). The concepts of gross and wasteful error are somewhat similar to the Type I (false positive) and Type II (false negative) errors used for statistical hypothesis testing. However, the terms are not used in standard statistical textbooks, including those dealing with classification error, and Altschul's definition therefore is specific to archaeological predictive modelling.

8.2.1 Gain and related measures

The most widely used method for model performance assessment is the calculation of the *gain* statistic of the model (Kvamme 1988b). Gain is calculated as follows:

$$G = 1 - \frac{p_a}{p_s}$$

where

pa = the area proportion of the zone of interest (usually the zone of high probability); and
ps = the proportion of sites found in the zone of interest.

If the area likely to contain sites in a region is small (the model is very precise), and the sites found in that area represent a large proportion of the total (the model is very accurate), then we will have a model with a high gain. Note that the gross error of a model is equal to $1 - p_s$. A similar measure, p_s/p_a or indicative value was used by Deeben *et al.* (1997) for the construction of the *Indicative Map of Archaeological Values* of the Netherlands (IKAW). The difference however is in the application of these measures. Whereas Kvamme's gain is exclusively applied to assess model performance, p_s/p_a has been used as a criterion for classifying individual variables into categories of low, medium and high probability.

Even though Kvamme ironically remarks that a model with a negative gain should result in firing the model developer, this is only true when testing the performance of the high probability zone – the low probability zone of course should have a low gain. In the Netherlands, performance assessment has usually been restricted to the calculation of the 'relative gain' p_s-p_a, with theoretical values that can range from 1 to -1 (Wansleeben and Verhart 1992). This measure however has the disadvantage of not specifying precision and accuracy. A 40% relative gain can be obtained by 80% of the sites contained in 40% of the area, but equally well by 50% of the sites contained in 10% of the area. The latter is of course more precise, but less accurate.

Other performance measures have been suggested, like Atwell-Fletcher weighting (Atwell and Fletcher 1985, 1987; Wansleeben and Verhart 1992; Kamermans and Rensink 1999). Verhagen and Berger (2001) pointed out that Atwell-Fletcher weighting is equivalent to normalizing p_s/p_a on a scale from 0 to 1, and it therefore belongs to the same category as Kvamme's gain. Wansleeben and Verhart (1992) developed the K_j-parameter:

$$K_j = \sqrt{p_s \frac{p_s - p_a}{p_w}}$$

where

$p_s > p_a$
p_w = the proportion of the area without sites.

K_j is a measure that favours accuracy over precision. The correction factor p_w was thought necessary because Kvamme's gain can never reach the maximum of 1, as the value of p_a in a model will never be equal to 0. There is therefore always a maximum gain value, dependent on the model itself. The parameter p_w can then be used as a maximum gain correction. However, this correction is easier to apply in a raster GIS-context, where all individual raster cells have equal sizes[51], and the number of 'non-site' cells is easily calculated, than in a vector-based model, where polygons may be of very different sizes, and polygons with no sites may even be absent. In order to obtain the maximum possible gain of a model, model optimisation methods have been developed (see 8.2.4).

[51] Obviously, the size of the raster cells will influence the value of p_w.

8.2.2 Measures of classification error

The issue of classification performance testing is extensively covered by Hand (1997). He notes that performance testing may be done for two reasons: to compare classifiers[52], and to obtain an absolute measure of performance. The most commonly used performance measure is the error rate, or rate of misclassification (*i.e.* the combination of gross and wasteful error). Hand points out that establishing the error rate, as the sole measure of performance, is not always what we are interested in. The different types of misclassifications may not be equally serious. This is exactly the case in predictive modelling, where gross error is usually considered to be more severe than wasteful error (Altschul 1988, 62-63).

Hand (1997) also offers a number of alternative measures of classification performance, in which he distinguishes four different flavours:

- *inaccuracy*, or the probability of misclassification – error rate is in this definition only seen as a *measure* of inaccuracy;
- *imprecision*, or the difference between an estimate and the true probability (*i.e.* accuracy in its statistical sense);
- *inseparability*, the similarity between different classes; and
- *resemblance*, the probability that a classification distinguishes between classes that are not there (Hand 1997, 110).

The model performance measures for predictive modelling discussed above only deal with the issue of inaccuracy. Imprecision does play a role when methods of model validation are discussed, but inseparability and resemblance have not featured as important issues in predictive modelling literature, with the exception of Rose and Altschul (1988).

Hand (1997, 100-102) objects to the calculation of error rate for measuring classification performance; it can produce a too optimistic assessment of the model's performance. He suggests using two other measures; the first one is called the *Brier* or *quadratic* score. This is basically the sum of squared deviations:

$$\frac{1}{n}\sum_{i-1}^{n}\sum_{j}\left(\delta(j \mid x_i) - \hat{f}(j \mid x_i)\right)^2$$

where

j = the class of interest
i =the object of interest
n = the number of objects

$\delta(j \mid x_i)$ = the classification of object i (1 if correct, 0 otherwise)

$\hat{f}(j \mid x_i)$ = the estimated probability that object i belongs to class j

The second one is called the *logarithmic score*, and is defined as follows:

$$-\frac{1}{n}\sum_{i-1}^{n}\sum_{j}\delta(j \mid x_i)\ln \hat{f}(j \mid x_i)$$

Both these measure will weigh the errors, taking care that classification errors that are far 'off the mark' are considered more serious than those that are found close to the class boundary. Of course Brier and logarithmic scores will result in figures that are not directly comparable to the error rate, or to each other. An additional feature is that they are equally applicable to multi-class cases, but they can also be used to calculate gross or

[52] The term *classifier* refers to the rule used for obtaining the classification.

wasteful error separately. From a cultural resource management perspective, the seriousness of a classification error does not really matter: a gross error is a gross error and will have the same consequences, whether it is found close to the class boundary or far off. Brier and logarithmic scores are therefore not suitable as absolute indicators of archaeological predictive model performance. They could however be used to judge if it is any use to change the class boundary. A model with many gross errors close to the class boundary between the site-likely and site-unlikely zones can be improved greatly by shifting the boundary slightly towards the site-unlikely class. With a model that contains many 'bad errors', substantially improving the accuracy might imply a dramatic decrease in precision (see also section 8.2.4 for more methods of model optimisation).

A statistical test, using the binomial distribution, can be used to establish confidence limits around a classification (Kvamme 1988b). As the classification of a predictive model that distinguishes between sites and non-sites is either right or wrong, the correctness of classification assignment represents a binomial distribution. The percent correct statistics can be considered as estimated mean probabilities of correct classification, and the corresponding confidence intervals can be calculated using the following equation:

$$\frac{p + \left(\frac{z^2_{\alpha/2}}{2n} \right) \pm z_{\alpha/2} \sqrt{\frac{p(1-p) + z^2_{\alpha/2}/4n}{n}}}{1 + z^2_{\alpha/2}/n}$$

where

p = proportion of correct predictions
n = the number of sites; and

$z^2_{\alpha/2}$ = the appropriate z-value at the 100(1-α) percent confidence interval for p.

This is a potentially important statistic, as it will not only give an estimate of the statistical precision of our proportion of correct predictions, but it also tells us where we can expect the proportion of correct predictions estimate to lie when an independent test set is used. When the proportion of correct predictions from a test set is outside the specified confidence limits, then we should be very suspicious of the quality of the original model - or of the test set data. Kvamme also notes that the confidence intervals obtained can be inserted into other formulas, like the gain statistic, in which p_s stands for the proportion of correct site prediction. It is therefore a potentially important test that can be used to combine performance measurement and statistical testing of a predictive model. However, I have not been able to track down a single case study for which it has been calculated.

Kvamme (1990) also introduced a more refined method of measuring model performance based on classification error measures. A 2x2 matrix is constructed, from which two performance statistics can be derived, the *conditional probability* and the *reverse conditional probability* (table 8.1-8.4; from Kvamme 1990). In a sample of site and non-site locations, the proportion of sites (P_s) is equal to 0.1, and the proportion of non-sites ($P_{s'}$) is 0.9. On the basis of the sample, a model is made resulting in two zones, a site-likely zone (M), taking up 26.5% of the sampled area and a site-unlikely zone (M'), covering 73.5% of it. The 10% site locations are divided as follows: 8.5% in the site-likely zone ($P_{s \cap m}$), and 1.5% in the site-unlikely zone ($P_{s \cap m'}$). The 90% percent non-site locations are divided as follows: 18% in the site-likely zone ($P_{s' \cap m}$) and 72% in the site-unlikely zone ($P_{s' \cap m'}$).

	M	M'	
S	$P_{s \cap m} = 0.085$ 'true positive'	$P_{s \cap m'} = 0.015$ 'false negative'	$P_{s} = 0.10$
S'	$P_{s' \cap m} = 0.18$ 'false positive'	$P_{s' \cap m'} = 0.72$ 'true negative'	$P_{s'} = 0.90$
	$P_{m} = 0.265$	$P_{m'} = 0.735$	

Table 8.1 The probabilities of percent correct assignment. M = model indicates sites, M' = model indicates no site. S = site observation, S' = non-site observation. Source: Kvamme (1990, 264).

By dividing the numbers in the matrix by the sum of the rows, the conditional probabilities are obtained (table 8.2). These are the standard measures of model accuracy: 85% of the sites are found in zone M (where they are predicted), and 80% of the non-sites are found in zone M' (where non-sites are predicted), with the corresponding classification errors of 15% and 20% respectively. Kvamme (1990) however states that is more important to know the probability of finding a site or non-site in each zone.

	M	M'	
S	$P_{m\|s} = 0.85$	$P_{m'\|s} = 0.15$	$P_{m\|s} + P_{m'\|s} = 1.00$
S'	$P_{m\|s'} = 0.20$	$P_{m'\|s'} = 0.80$	$P_{m\|s'} + P_{m'\|s'} = 1.00$

Table 8.2 The conditional probabilities obtained from table 8.1.

This can be established by calculating the reverse conditional probabilities, which is done by dividing the numbers in the matrix by the sum of the columns (table 8.3). The probability of site presence in the site likely zone is 0.32, whereas it is only 0.02 in the site-unlikely zone.

	M	M'
S	$P_{s\|m} = 0.32$	$P_{s\|m'} = 0.02$
S'	$P_{s'\|m} = 0.68$	$P_{s'\|m'} = 0.98$
	$P_{s\|m} + P_{s'\|m} = 1.00$	$P_{s\|m'} + P_{s'\|m} = 1.00$

Table 8.3 The reverse conditional probabilities obtained from table 8.1.

The performance of the model is then determined by comparing these probabilities to a by-chance model. In order to do this, the reverse conditional probabilities are divided by the by-chance (a priori) probabilities, taken from the right-hand column in table 8.1.

$P_{s|m}/P_{s}$, for example, can then be translated as a '3.2 times better performance in site prediction than a by-chance model'. Incidentally, this figure equates to the indicative value (p_s/p_a) used by Deeben *et al.* (1997). The innovation is found in the other statistics, especially the performance for false negative and false positive outcomes. Obviously, this approach will only work when a model is made based on site and non-site data.

	M	M'
S	$P_{s\|m}/P_s = 3.2$ 'true positive'	$P_{s\|m'}/P_s = 0.2$ 'false negative'
S'	$P_{s'\|m}/P_{s'} = 0.76$ 'false positive'	$P_{s'\|m'}/P_{s'} = 1.09$ 'true negative'

Table 8.4 Performance statistics for the model from table 8.1.

In a later paper, Kvamme (1992, 14) stated that obtaining non-site data for model development and testing purposes from the surveyed zones is not to be advised, as these non-site locations will be close to site locations and therefore spatially auto-correlated with them (see section 8.2.8). The site-unlikely model (M') developed with non-site locations from the surveyed zones will in those cases be very similar to a site-likely model (M). Gibbon *et al.* (2002) reached similar conclusions when trying to use non-site data from surveyed areas for model development. In order to overcome the problem, Kvamme proposes instead to sample the non-site locations from the background environment (*i.e.* the whole study area), and defends this by pointing out that sites usually occupy only a small proportion of the total area. The probability that a sample of non-sites taken from the background area actually contains a site is therefore very low, and will not drastically influence the outcome of the model (see also Kvamme 1988b, 402). The argument is not wholly convincing; when background non-sites are used for model development, the site-unlikely model (M') will of course be very similar to a by-chance model. In fact, the whole procedure of taking separate background non-site samples then becomes unnecessary. The only thing that is needed is an estimation of the site to non-site ratio, which will come from the surveyed zones.

8.2.3 Performance optimisation methods

In order to produce a model with maximum performance, optimisation methods have been developed to be used during model building. The best known of these, the *intersection method*, was introduced by Kvamme (1988b). It has especially been applied with logistic regression modelling (see *e.g.* Warren 1990a; 1990b; Carmichael 1990; Warren and Asch 2000; Duncan and Beckman 2000), and is restricted to models that use a site/non-site approach. As these models result in a probability map of site occurrence per grid cell, it is very easy to reclassify the map in, *e.g.*, 10 categories of site probability, and compare these to the actual percentage of sites and non-sites contained in each probability zone. In this way, cumulative curves can be constructed of both site and non-site prediction accuracy. At the intersection point of the curves, the probability of gross error is equal to the probability of wasteful error. However, as Kvamme (1988b) notes, a trade-off could be made, sacrificing precision for greater accuracy if we want to decrease the gross error. The method is therefore very useful as a tool for the final classification of the model into zones of low and high probability, and provides an easily interpretable tool for making the trade-off between model accuracy and precision.

The K_j-measure (Wansleeben and Verhart 1992) was originally developed in the context of finding the optimum performance of a single-variable model as well. By calculating K_j for each individual category, the 'most successful' category can be found. This category is then included in the model, and K_j is calculated again for the remaining categories. This procedure is repeated until all categories are included in the model. At each step, the relative gain (or any other gain measure) can be used as a measure of the performance of the model, and the cut-off point between high and low probability can be chosen on the basis of it. Verhagen and Berger (2001) took the step of combining the individual rankings into gain development graphs, which can easily be used to decide where to place the boundary between low, medium and high probability zones. Warren and Asch (2000) have published similar curves to be used with logistic regression models.

Whitley (2005a) suggests to create multiple deductive models with a number of different modelling criteria, and compare these to a test data set in order to see which model best fits the data. This is an additional method of model optimisation, well suited for developing the best possible deductive hypothesis of site location.

8.2.4 Performance assessment of dutch predictive models

How well do currently used Dutch predictive models perform? A judgement of their performance can only be made using gain or gain-like measures, as the site/non-site approach to predictive modelling has never been used in the Netherlands. Furthermore, many predictive maps made in the Netherlands are not based on quantitative analysis, but on expert judgement, and do not provide sufficient information to judge both their accuracy and precision. The exception to the rule is the IKAW (Deeben *et al.* 1997) that was developed using the indicative value to decide whether a zone should be low, medium or high probability. In the papers published by Deeben *et al.* (1997; 2002), the performance of the predictive map is judged by comparing the values of p_a and p_s, without actually proceeding to presenting a measure of (relative) gain. This can however be easily calculated from the figures given, which are only provided for one of the thirteen archaeo-regions analysed. For this particular region, the Eastern Sandy Area, the figures presented in Deeben *et al.* (2002) are as follows (table 8.5):

	p_a	p_s	Kvamme's gain
low probability	0.629	0.249	-1.522
middle probability	0.193	0.222	0.130
high probability	0.178	0.529	0.663

Table 8.5 Performance statistics for the IKAW (Eastern Sandy Area).

From these figures it can be concluded that the model is not extremely accurate; only 52.9% of the known sites is captured in the high probability zone. However, it is relatively precise, as the high probability zone only takes up 17.8% of the area, and therefore the resulting gain of the high probability zone is quite high. When analysing the 2nd version of the IKAW (Deeben *et al.* 2002), substantial differences between regions can be observed. Kvamme's gain values for the high probability zone vary between 0.20 (the loess area of Limburg) and 0.79 (the clay area of Zeeland) (figure 8.2). Whereas the 5 sandy Pleistocene areas, together taking up 51.1% of the Netherlands, exhibit very similar accuracy, precision and Kvamme's gain values, the other areas show substantial differences in performance. Poor performance is observed for the loess area of Limburg and the dune area of Holland, whereas very good performance is found for the peat area of Friesland and clay area of Zeeland. These last two however 'pay' for their high gain values with very low accuracy.

The IKAW is a quantitative model; RAAP Archeologisch Adviesbureau has produced expert judgement predictive maps for a number of years (see van Leusen *et al.* 2005). How well do these perform? Not all of them have been compared quantitatively to the distribution of archaeological sites, but two reports do provide such figures. The first one is the predictive map of the Roman *limes* in the province of Gelderland (Heunks *et al.* 2003), which is primarily based on the distribution of fossil river channels dating from the Roman period. The figures presented in the report are as follows (table 8.6):

	p_a	p_s	Kvamme's gain
low probability	0.387	0.048	-6.590
medium probability	0.394	0.520	0.156
high probability	0.219	0.378	0.180

Table 8.6 Performance statistics for the limes map of Gelderland (Heunks *et al.* 2003)

The sites involved for performance testing are all dated to the Roman period. Evidently, the map is not accurate in a quantitative sense. The model seems to be very good at explaining where *no* sites will be found.

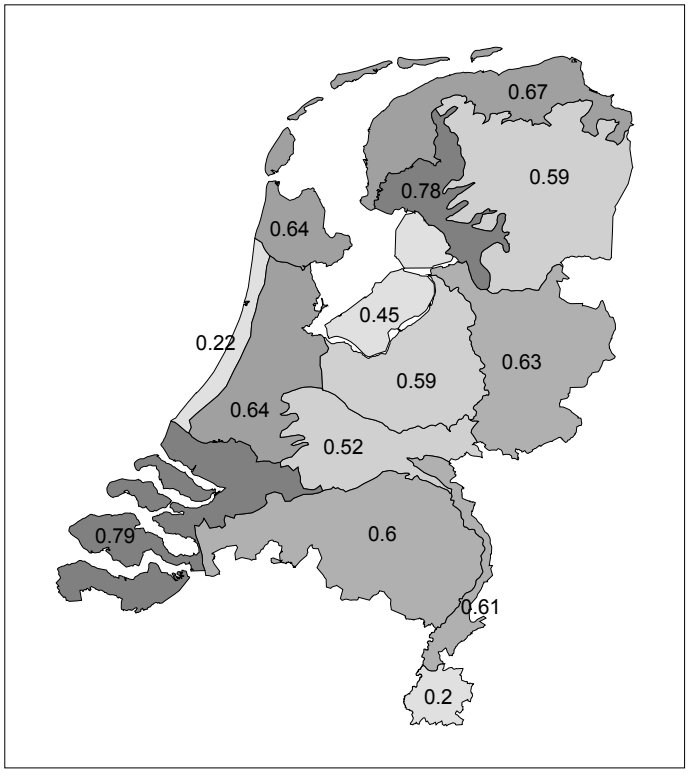

Figure 8.2 Kvamme's gain values per archaeo-region for the 2nd generation IKAW.

The predictive map for the municipality of Ede, province of Gelderland (Heunks, 2001) shows quite a different picture (table 8.7).

This model is very accurate, capturing 79.4% of the sites in the high potential zone. However, it is not terribly precise, resulting in a Kvamme's gain value of 0.411. The expert judgment maps were not optimised in a quantitative sense, which is reflected in the gain values calculated. In the case of the limes predictive model however, a quantitative optimisation would not have resulted in very accurate or precise outcomes either, as the sites seem to be relatively evenly distributed on the various landscape units distinguished. A large amount of sites is found in the medium probability zone, and it has to be pointed out that this refers to sites that have been checked in a desktop study for their location and content. Heunks *et al.* (2003) explain that the unit of riverbank

deposits (13.4% of the area with 23.9% of the sites) had been classified in the medium probability zone because it in fact consists of two zones, one of low and one of high probability, that could not be separated on the basis of the geological maps used. Evidently, the model could have been made more accurate by including these riverbank deposits into the high probability zone, but even then the gain of the model would not have been very high.

	p_a	p_s	Kvamme's gain
low probability	0.298	0.138	-1.166
medium probability	0.235	0.069	-2.419
high probability	0.467	0.794	0.411

Table 8.7 Performance statistics for the municipal map of Ede (Heunks 2001)

8.2.5 Comparing classifications

When comparing models, our main interest lies in knowing whether there is a difference between the original model and the new model, irrespective of whether this is based on new data or on data that was kept behind as a test set. One simple method to obtain information on the difference between two classifications is the calculation of the kappa coefficient.

In order to quantify the difference between two classifications, an error matrix of the classifications in both models is made. Especially when dealing with raster GIS-models, we can calculate the number of grid cells that were classified differently in each model. From the error matrix a measure of model stability, the kappa coefficient, can be calculated. The kappa coefficient was developed by Cohen (1960) to measure the degree of agreement between models, after chance agreements have been removed. It was recommended as standard procedure for measuring remote sensing classification accuracy by Rosenfield and Fitzpatrick-Lins (1986), and has been used by Hobbs *et al*. (2002) to measure the stability of the MnModel. Kappa is calculated as follows (Banko 1998)[53]:

$$K = \frac{p_o - p_e}{1 - p_e}$$

where

p_o = observed agreement
p_e = expected agreement

The expected agreement is the 'by chance' probability of agreement between the models. In the case of a classification into three categories, the probability of agreement between the models is 1/3. The probability of each grid cell combination of the models in a by chance agreement is 1/9; there are 3 possible agreement combinations, so the overall agreement probability is 1/3.

In order to obtain the value of p_o, the actual proportion of cells in agreement is calculated. The resulting value of kappa will be a number between -1 (perfect disagreement) and 1 (perfect agreement). The variance of kappa can be also calculated to test whether the classification accuracy differs significantly from chance agreement. However, this calculation is much more complex (Hudson and Ramm, 1987):

[53] Calculation of the kappa coefficient, but not its variance, is included as a standard tool in Idrisi.

$$\sigma^2[\mathrm{K}] = \frac{1}{N}\left[\frac{\theta_1(1-\theta_1)}{(1-\theta_2)^2} + \frac{2(1-\theta_1)(2\theta_1\theta_2-\theta_3)}{(1-\theta_2)^3} + \frac{(1-\theta_1)^3(\theta_4-4\theta_2^2)}{(1-\theta_2)^4}\right]$$

where

$$\theta_1 = \sum_{i-1}^{r}\frac{X_{ii}}{N}$$

$$\theta_2 = \sum_{i=1}^{r}\frac{X_{i+}X_{+i}}{N^2}$$

$$\theta_3 = \sum_{i-1}^{r}\frac{X_{ii}(X_{i+}+X_{+i})}{N^2}$$

$$\theta_4 = \sum_{i=1,j=1}^{r}\frac{X_{ij}(X_{i+}+X_{+i})^2}{N^3}$$

X_{ii} is the count in row i and column i (*i.e.* the cells that are in agreement); X_{i+} is the sum of row i, and X_{+i} the sum of column i. Significance can then be tested by calculating

$$Z = \frac{\mathrm{K}_1 - \mathrm{K}_2}{\sqrt{\sigma_1 - \sigma_2}}$$

The kappa coefficient can be used to measure the difference between a classification and a random classification, and between two classifications.

8.2.6 Comparing classifications: an example from practice

In 2003, RAAP Archeologisch Adviesbureau produced a revised version of the IKAW for the province of Gelderland[54]. For financial reasons, the provincial authorities did not commission a complete revision of the base maps (*i.e.* the soil maps and palaeo-geographic maps of the area), but only a reclassification of the units used for the IKAW, as it was felt that the IKAW contained too many classification errors. The archaeological data set was thoroughly screened and revised, and the reclassification was done using expert judgment instead of quantitative optimisation. At the time, the resulting map was not compared quantitatively to the IKAW, and no report has appeared to justify the reclassification made. However, since the base material was identical to that used for the IKAW (2nd generation; Deeben *et al*. 2002), it is very easy to compare the classifications (table 8.8; figure 8.3).

[54] As part of the map of cultural historical values of Gelderland (Gelderse Cultuurhistorische Waardenkaart, or GCHW).

	high	medium	low	total
IKAW (km^2)	1203.99	1035.2	2538.96	4778.15
%	25.20%	21.67%	53.14%	
GCHW (km^2)	1121.49	1504.61	2138.31	4764.41
%	23.54%	31.58%	44.88%	
IKAW, number of sites	3157	1831	1430	6418
%	49.19%	28.53%	22.28%	
GCHW, number of sites	3144	2269	981	6394
%	49.17%	35.49%	15.34%	
IKAW				
Kvamme's gain	0.49	0.24	-1.38	
relative gain	23.99%	6.86%	-30.86%	
p_s/p_a	1.95	1.32	0.42	
GCHW				
Kvamme's gain	0.52	0.11	-1.93	
relative gain	25.63%	3.91%	-29.54%	
p_s/p_a	2.09	1.12	0.34	
IKAW, no. of observations	4458	2517	1914	8889
%	50.15%	28.32%	21.53%	
GCHW, no. of observations	4330	3141	1391	8862
%	48.86%	35.44%	15.70%	
IKAW				
Kvamme's gain	0.50	0.23	-1.47	
relative gain	24.95%	6.65%	-31.60%	
p_s/p_a	1.99	1.31	0.41	
GCHW				
Kvamme's gain	0.52	0.11	-1.86	
relative gain	25.32%	3.86%	-29.18%	
p_s/p_a	2.08	1.12	0.35	

Table 8.8 Comparison of performance of the IKAW and the GCHW-maps. The comparison is made using both the screened archaeological data-set (number of sites) and the original ARCHIS observations.

The reclassification resulted in a smaller area of low probability, with a lower gain. The area of medium probability is substantially increased, but its gain is smaller, so it better complies with the demand of neutral predictive power. For the high probability zone, very little changed, although some improvement in gain can be observed. Even though the archaeological data set was screened, and reduced by about 28%, the performance of the models based on the screened data set is not very different from the old data set. The shifts in classifications are given in table 8.9 (the totals are slightly different because of the occurrence of no data zones, that are also defined differently).

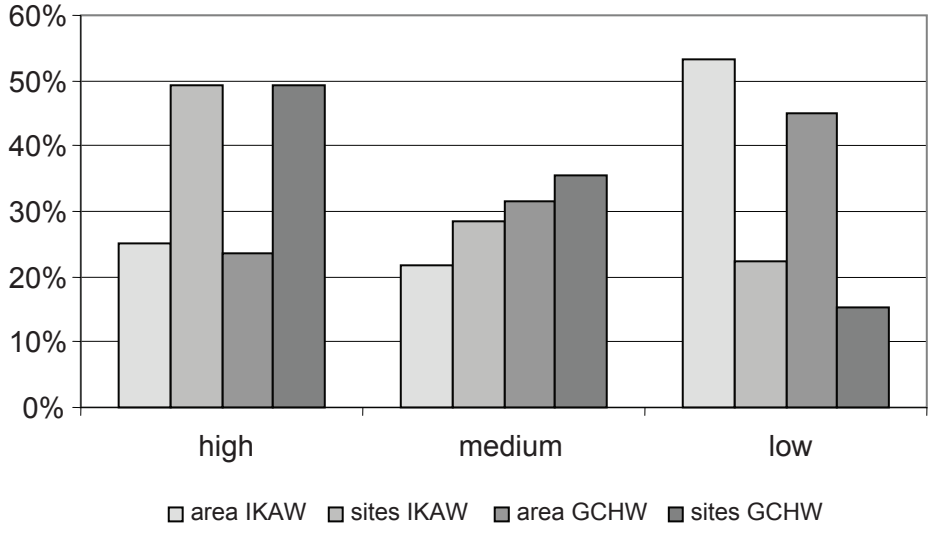

Figure 8.3 Comparison of the IKAW and GCHW maps, based on the screened archaeological data set.

The kappa-coefficient is 0.498, which points to a moderate difference between the classifications (66.5% of the area retains its original classification). When looking at the shifts between categories, the high and low probability categories seem to remain relatively stable (resp. 74.0% and 70.0% keep the IKAW-classification). In contrast, only 49.4% of the medium probability area is retained, and 33.6% of it is reclassified as low probability. In absolute figures however, the 29.3% of the low probability area reclassified as medium probability contributes more to the total area of medium probability, leading to a larger medium probability zone than before. More interesting is the fact that 41.2% of the sites initially found in the low probability zone are now contained in the medium probability zone, whereas only 8.7% of the sites originally located in medium probability are now found in the low probability zone. The main effect of reshuffling the categories has been that the number of sites contained in low probability zones has been reduced by 31.0%. This points to the fact that the archaeologist responsible for the reclassification primarily aimed at reducing the gross error of the model.

km^2	GCHW-high	GCHW-medium	GCHW-low	Total IKAW
IKAW-high	888.031	265.054	46.9223	1200.007
IKAW-medium	175.631	510.888	347.762	1034.281
IKAW-low	17.5613	727.929	1741.27	2486.76
Total GCHW	1081.223	1503.871	2135.9543	4721.049
Shift IKAW-GCHW in %area	GCHW-high	GCHW-medium	GCHW-low	
IKAW-high	74.00%	22.09%	3.91%	
IKAW-medium	16.98%	49.40%	33.62%	
IKAW-low	0.71%	29.27%	70.02%	
number of sites	GCHW-high	GCHW-medium	GCHW-low	Total IKAW
IKAW-high	2662	426	53	3141
IKAW-medium	413	1257	160	183
IKAW-low	69	586	768	1423
Total GCHW	3144	2269	981	6394
Shift IKAW-GCHW in %sites	GCHW-high	GCHW-medium	GCHW-low	
IKAW-high	84.75%	13.56%	1.69%	
IKAW-medium	22.57%	68.69%	8.74%	
IKAW-low	4.85%	41.18%	53.97%	

Table 8.9 Comparison of IKAW- and GCHW-classifications.

8.2.7 Spatial autocorrelation and spatial association

Spatial autocorrelation refers to the fact that objects that are close together, tend to have similar characteristics. Whitley (2005b) distinguishes between first-order autocorrelation (*i.e.* that archaeological sites tend to be located close together) and second-order autocorrelation - the first-order autocorrelation is influenced by variables that (unconsciously) determine spatial decisions, and these are the very variables that we use for predictive modelling. The issue has not been studied in detail by archaeologists, and there is no consensus on whether it is something that always needs to be avoided or that can be used to advantage in predictive modelling. Especially second-order spatial autocorrelation seems at first sight advantageous to predictive modellers, as it may be a strong predictor of site occurrence (Millard 2005). And in fact, spatial autocorrelation (and anti-correlation!) is used extensively for prediction and interpolation purposes in the field of geo-statistics. It is however clear that first-order spatial autocorrelation has a strong effect on the outcome of significance tests (Kvamme 1993), and Millard (2005) points out that neglecting the effects of spatial autocorrelation in archaeological predictive modelling leads to an overestimation of the predictive power of the models (see also van Leusen *et al.* 2005: 67-68). The use of spatially auto-correlated data sets for statistical inference will result in an overestimation of statistical significance, leading to inflated performance statistics and narrower confidence intervals. It is therefore advisable to correct for spatial autocorrelation in the archaeological site data. In order to measure spatial autocorrelation between objects, two indices are often used, Moran's I[55] and Geary's c[56]. Kvamme (1993) suggests a method using Moran's I for calculating the 'effective sample size', in order to correct the over-optimistic estimates that will be obtained using auto-correlated data. This is a straightforward and useful technique to prevent spatial autocorrelation influencing both the construction of a predictive model, as well as the measurement of its performance.

One way to use spatial autocorrelation for model testing is as a means to detect 'outliers' in a model. If the residuals of a logistic regression model exhibit spatial autocorrelation, then we can be reasonably confident that one or more explanatory variables are missing in the model. However, as far as can be judged, an analysis of the residuals of logistic regression models has never been published, at least not in 'white' literature.

Spatial association (also known as *spatial cross-correlation*) is the amount of correlation between two spatial variables. Spatial association should be analysed in the preliminary stages of model building, where correlation between the input variables for a model needs to be checked. A simple procedure to detect spatial association by calculating Moran's I on the covariance between two maps is described by Kvamme (1993), but other measures are available as well, like the one developed by Goodchild (1986)[57]. Bonham-Carter (1994) suggests a technique that can be used with predictive models that are based on site density transfer mapping, and that use more than one variable. The ratio of observed to predicted sites (the indicative value) per area unit in the model can in those cases serve as a measure of violation of the assumption of conditional independence of the variables, as the sum of the predicted probabilities per area unit should equal the number of observed sites in the case of complete independence.

The use of spatially correlated data sets for constructing predictive models of any kind should be strongly discouraged, as it will have an effect on the outcome of statistical significance testing. Millard (2005) points out that logistic regression models do not provide a safeguard against spatial association, and may therefore produce apparently statistically valid correlations while in fact using spatially associated datasets.

8.2.8 Summary and discussion

In the current section I have so far reviewed criteria and measures for the assessment of model performance (8.2.2 and 8.2.3), and measures and methods for optimising model performance (8.2.4 and 8.2.6), but have refrained from discussing their utility for the task in hand – namely, to give an adequate description of predictive

[55] Included as a standard tool in Idrisi and ARC/INFO GRID.

[56] Included as a standard tool in ARC/INFO GRID.

[57] Incorporated in ARC/INFO GRID as a standard tool.

model quality that can be used in a meaningful way in a cultural resource management context. Here I will highlight the problems with using the existing measures for model performance assessment, and present my views on how to proceed.

Performance assessment of archaeological predictive models can be done using two approaches: the calculation of gain and gain-like measures, and the calculation of classification error. The latter approach has, to date, not been used for model performance assessment in the Netherlands, as Dutch predictive modellers have not adopted the American site/non-site approach; this is because of a lack of controlled survey data available for model construction (see section 8.4; Brandt *et al.* 1992). Almost all models published up to date in the Netherlands have adopted a three-zone classification of high, medium and low probability of site occurrence, using 'site density transfer' mapping. These zones are defined either intuitively or by relatively simple weighted overlay mapping[58]. In order to calculate classification error, classes must be mutually exclusive: a site cannot be a non-site, and if one is found in the 'non-site' zone, then we have a classification error. If a classification only establishes zones of relative density, as with Dutch predictive maps, we cannot say that any single site is incorrectly classified and classification error methods then cannot be used. For the same reason, the intersection method of model optimisation for trading off accuracy and precision has not been applied in Dutch predictive modelling either. However, performance optimisation is inherent to the procedures followed for building the IKAW model: Deeben *et al.* (1997) used cumulative curves of site proportions to decide where to place the boundary between low, medium and high probability.

Kvamme's gain is the only measure that can easily be transferred to Dutch predictive modelling practice, and in fact the alternative performance measures that have been suggested by Dutch predictive modellers, like the 'indicative value', Atwell-Fletcher weighting, relative gain and K_j are all very similar in nature to Kvamme's gain.

Some authors have criticized the use of gain measures for performance assessment because of the inbuilt assumption of comparison to a 'by chance' model (Whitley 2005b). They claim that a model performing better than a 'by chance' model is nothing to be proud of, and could easily be made using the wrong modelling assumptions and parameters. While this is true, from a cultural resource management perspective a model should be accurate and precise in the sense used in predictive modelling. The 'by chance' model is the worst performing model imaginable, and it therefore makes sense to calculate performance measures this way.

Gain combines the two performance criteria of accuracy and precision in one, easily calculated measure. However, it does not fully cover the issue of performance assessment. 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 (model A), or by including 80% of the sites in 40% of the area (model B; see table 8.10). In model A, the risk of encountering a site in the low probability zone is greater than in model B, which is reflected in Kvamme's gain values of resp. -0.75 and –2.0 for the low probability zone. An assessment of model quality should therefore include performance measures for the low probability or site-unlikely zone as well, and preferably a comparison measure of the two zones as well. This is easily done using the ratio of the indicative value (p_s/p_a) for each probability zone. For model A, the ratio of indicative values of the high and low probability zone is equal to 2.0/0.57 = 3.5[59]; for model B, this ratio is 2.0/0.33 = 6.0, indicating a better performance for model B. However, even when using all these three measures it may be still be difficult to decide what is the best performing model. This is illustrated by the fact that a ratio of indicative values of 6.0 can also be obtained by model C, containing 90% of the sites in 60% of the area; this model has a Kvamme's gain of 0.33 for the high probability zone, and of –3.0 for the low probability zone. Intuitively, one would judge this model to be performing worse than model B because of the low gain in the high probability zone, and the lower relative gain of 30% instead of 40%. But in fact, it may be a very good model for spatial planning purposes, as its low

[58] Incidentally, this type of predictive modeling is not absent in North-American predictive modeling, but it does not figure as prominently in literature on the subject (with the notable exception of Dalla Bona 1994; 2000). Even intuitive models are used in the United States (see *e.g.* Griffin and Churchill 2000).

[59] This equates to stating that in model A, the probability of finding a site in the high potential zone is 3.5 times higher than in the low potential.

probability zone has a very low probability of encountering sites and greatly reduces the archaeological risk in 40% of the area.

The use of medium probability zones poses an additional problem for model performance assessment. Because these are zones of no predictive power, they mainly serve to minimize the zones of high and low probability. The gain of the high and low probability zone will then always be inflated, and will not give a good impression of the performance of the whole model – in the end, we are not particularly interested in a model where the majority of the study area is medium probability. Depending on whether we want to emphasize accuracy or precision, the medium probability zone should be included in the high or low probability zone for model performance assessment purposes. For the Eastern Sandy Area of the IKAW the calculated gain of 0.663 (see section 8.2.5) for the high probability zones becomes a gain of 0.506 when including the medium probability zone.

	p_s(high)	p_a(high)	p_s(low)	p_a(low)	Kvamme's gain	indicative value	ratio i.v.
Model A	0.6	0.3	0.4	0.7	0.5 -0.75	2.0 0.57	3.5
Model B	0.8	0.4	0.2	0.6	0.5 -2.0	2.0 0.33	6.0
Model C	0.9	0.6	0.1	0.4	0.33 -3.0	1.5 0.25	6.0

Table 8.10 Example of different performance characteristics of three hypothetical predictive models. Model B performs better while having the same gain as model A for the high probability zone. However, model C may be the best for spatial planning purposes.

The issue of defining model performance goals has rarely featured in predictive modelling literature, although some exceptions are found. The state-wide predictive model of Minnesota (MnModel) for example was to capture 85% of all sites in no more than 33% of the area, equating to a Kvamme's gain value of 0.61 (Hobbs 2003). Gibson (2005) indicates that a 'good working model' should have at least 70% of all sites in no more than 10% of the area, resulting in a Kvamme's gain value of 0.86 or more, which is a very high standard of performance. It can however be doubted if very high gains are attainable goals for many predictive models. Ducke and Münch (2005) believe that gain values of around 0.5 may be very typical for European predictive models. Ebert (2000) states that the 'reported accuracies of inductive predictive modelling seem to hover in the 60-70% range'. Assuming that he refers to accuracy in the sense that it has been used above, this means that the high probability zones of predictive models never capture more than 70% of the site observations. This relatively low accuracy of many predictive models may partly be due to the model performance optimisation methods used and to the lack of predefined goals for performance; especially the intersection method is meant to offer the ideal compromise between the demands of accuracy and precision. As Kvamme (1988b) pointed out, accuracy may be increased with this method, but only at the cost of decreasing precision. The underlying problem therefore is that many models are not precise enough, and Ebert pessimistically concludes that they will not become any better.

From the point of view of protection of the archaeological heritage, accuracy is a much more important characteristic of a predictive model than precision. Low accuracy implies a high archaeological risk, because any area that is classified into the low probability or site-unlikely category will be threatened more easily. Spatial planners will feel that these areas can be developed with less risk, and will tend to have a preference for low probability instead of high probability zones. Furthermore, in most cases there will be no obligation to do survey in these areas. This means that the less accurate a model is, the higher the archaeological risk will be in

the low probability zones. In establishing criteria for model quality, accuracy should therefore come first, and precision second. In this way, it is also much easier to compare the performance of predictive models. By fixing the desired accuracy of a model to a predefined level, models can only 'compete' on precision. It is then very easy to decide which model performs best, and the, sometimes confusing gain measures are no longer necessary for performance assessment. However, it also means that we sometimes will have to content ourselves with a model that is not terribly precise.

In everyday practice, archaeologists working in archaeological heritage management are not overly concerned with quantitative quality norms for predictive models. They usually equate a predictive model to a theory of site location preferences, not to a statistical model, and this is what they expect a predictive map to depict. A very simple example is found in the fact that in the Dutch coastal and fluvial areas, it is essential to be able to distinguish between the uninhabitable, submerged zones, and the inhabitable, dry zones. These zones have shifted during prehistory, and can only be recognized on the basis of lithological and pedological characteristics of the soil. This is considered a predictive model: a binary division between zones of interest, and zones of no interest, based on recognizable characteristics of the (palaeo-)landscape. A third 'medium probability' zone, while sometimes recognized, mainly indicates that there is not enough information available to distinguish between the former two, or that we are dealing with a transition zone, where *e.g.* off-site phenomena may be found. The predictive map will then lead to a more specific archaeological question: if this zone was habitable, what types of sites can we expect? It is this question that will in the end determine the survey strategy to be used for detecting archaeological sites.

Obviously, with this approach to predictive modelling it is impossible to impose performance criteria on the models. We cannot artificially reduce the area taken up by *e.g.* a fossil creek bed to increase precision, nor can we demand that these creek beds should contain 85% of all known sites. On the other hand, it is possible to calculate performance measures for these expert judgement models from the available archaeological data sets. This should be an obligatory step after the construction of expert judgment models. After all, we need criteria to decide whether zone A is more important than zone B; to decide whether model A is better than model B; and to decide whether we need additional information to reduce uncertainty. Without a quantitative basis, these decisions will remain the province of the archaeological experts, whose knowledge cannot be checked against independent evidence.

8.3 Validation of model performance

Validation, as defined by Rose and Altschul (1988), involves verifying a model's performance on 'independent data, on part of the sample that was excluded from the model-building process, or on internal criteria'. Validation in this sense is used for determining the classification error of a model, and compares the classification error obtained from the design data set with a test data set. When calculating the classification error directly from the data set used for designing the model, we can expect the resulting 'apparent' or 'resubstitution error rate' (Hand 1997, 121) to present a biased and optimistic estimate of the true or actual error rate, *i.e.* the error rate that would be obtained from the total population. This is especially true with small to moderate-sized data sets. The reason for this is, that the classification rule is optimised to the design set. New data (assuming that it will come from the same population) will usually have a slightly different distribution, and therefore should be expected to show a larger error rate. We should therefore always expect validation to exhibit larger errors for an independent data set than for the design set. However, this does not tell us whether we should reject or accept the model. The only thing validation will do is give a more realistic indication of model performance than is obtained by calculating performance statistics using the design data set itself. This in turn implies that performance statistics should always be calculated using an independent data set. This 'external' or 'double' validation has not always proved possible in predictive modelling, so several techniques have been developed for 'internal' or 'simple' validation. These are described in section 8.3.1. However, the procedures can equally well be used with independent data sets. In section 8.3.2 the utility of validation methods for assessing predictive model performance will be discussed.

8.3.1 Simple validation techniques

Both Rose and Altschul (1988) and Kvamme (1988b; 1990) have discussed several methods for what they call simple validation, and what is also known as internal testing. A clear distinction should be made between *split sampling* methods on the one hand, that keep data from the available sample apart to see whether the model is any good at predicting the data that is left out from model building, and methods that re-use parts of the complete data set in order to obtain a more accurate model, like jackknife sampling. These methods are also known as *resampling* techniques. Split sampling is a classical validation method, as it uses a second data set for testing, and the procedures for validation by means of split sampling are simple and equal to those used for external or double validation. Resampling is a way to improve the performance of the model by artificially increasing the sample size of the design data set. Despite the difference in application of split sampling and resampling, I have decided to describe these techniques together in this section, as they are closely connected in a technical sense, and have been discussed together in other publications as well.

Split sampling requires randomly splitting the sample in two equal parts, building the model with one half, and validating it with the other half. A disadvantage of this method is that it severely reduces the data set available for model building. As a rule of thumb, the data set should not be split in half unless the total sample size is greater than 2p+25, where p is the number of parameters in the model (such as distance to water; Rose and Altschul 1988). It can easily be applied to establish if there is a difference between the model and the data that was kept behind, using all types of performance measures and statistical estimates, but in practice it has only been used to compare classification error rates, as discussed in 8.2.3.

The simplest resampling method available is *cross-validation*[60]. It refers to dividing the sample into a number of randomly chosen, roughly equal sized subsets (this is also known as *rotation*; Hand 1997, 122). 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.

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 comes very close to what is generally known as *jackknife sampling*. However, jackknife error estimation deals differently with establishing the error rate (Hand 1997). 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, 122). The error rate is determined at each analysis by using the complete data set (which of course contains no double observations). Improvements of the bootstrap error rate calculation have resulted in a measure known as the *.632 bootstrap* (Efron and Tibshirani 1993), which is the most accurate error estimator that has been developed up to date. Current statistical opinion therefore favours this error measure as the method of choice (Hand 1997), and jackknife sampling is considered by some to be of largely historical interest (Mooney and Duval 1993).

Table 8.11 summarizes the difference between the methods in terms of the sampling and analysis strategy applied. As computer power has increased enormously, bootstrap and jackknife methods are now much easier to implement than before, and are therefore gaining rapidly in popularity, especially since they are thought to perform better in estimating error rate. The differences in error determination between traditional cross-validation on the one hand, and jackknife and bootstrap sampling on the other hand are however not so easily explained, as this depends on quite complex statistical reasoning. Efron and Tibshirani (1993) and Hand (1997) provide more information on this subject.

[60] 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.

split sampling (hold-out method)	keeps a test set apart, usually half of the data determines error rate with the test set error rate of test set and original data set are compared, not averaged
cross-validation	divides sample randomly into k subsets withholds each subset from analysis in turn constructs k models with remainder of data and determines k error rates using withheld data total error rate is estimated by averaging error rates across models
leave-one-out (LOO)	same as cross-validation, but k = n (1 observation left out at a time)
jackknife	same as LOO, but error rate determined differently
bootstrap	takes a random sample with replacement of size n k times determines the error rate using the original data set total error rate is estimated by averaging error rates across models extended error rate estimators have been developed

Table 8.11 Different internal validation methods compared.

Unfortunately, no single definition of resampling can be found in statistical literature. In this section, I have decided to group all techniques that reuse the design data set under resampling. However, Simon (1998) specifically excludes cross-validation and jackknife sampling from resampling, as the former are methods that systematically exclude observations from the data set. Every jackknife analysis will therefore produce the same result with a given data set. Simon also specifies permutation resampling (*i.e.* without replacement) as a separate technique. This is also known as randomisation. It should also be noted that the formal definition of bootstrap resampling as given by Simon (1998) is less narrow than it is given in table 8.11: in fact, resampling can be carried out with subsets of any size. Resampling is also closely related to Monte Carlo-simulation[61], with the difference that Monte Carlo-simulation does not use sample data, but creates 'virtual' samples instead. Table 8.12 summarizes the main differences between the various methods.

jackknife sampling / cross-validation	systematically excludes observations number of simulations is equal to number of subsets
bootstrap resampling	randomly excludes observations number of simulations unlimited resamples with replacement
permutation resampling	as bootstrap, but resamples without replacement
Monte Carlo-simulation	number of simulations unlimited only uses 'virtual' data

Table 8.12 Resampling techniques available for statistical inference.

Resampling is currently positioned as an alternative to classical statistical inference by some authors (Simon 1997; Lunneborg 2000). In fact, both Simon (1969) and Efron (1979) developed bootstrapping specifically for this purpose. More traditional statisticians however only resort to bootstrapping in cases where classical inferential solutions are not available. Lunneborg (2000) mentions a number of limitations of classical statistical (parametric) inference. Especially small sample size, small population size and the assumption of random

[61] Simon (1997) even classifies resampling as a subset of Monte Carlo-methods.

sampling are limiting 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. Resampling however does need representative samples just like parametric techniques: a biased sample will also produce biased results with resampling. Simon (1997) adds that resampling is a more intuitive way of dealing with statistical inference, and consistently leads to better statistical problem-solving by students and non-statisticians than the use of classical parametric methods. A good and accessible overview of the discussion on resampling can be found on http://seamonkey.ed.asu.edu/~alex/teaching/WBI/resampling.html.

8.3.2 Simple validation and predictive modelling

Both Rose and Altschul (1988) and Kvamme (1988b; 1990) have used jackknife sampling as a method to develop a 'composite model' of all possible models that can be made by leaving out one observation at a time. In their approach, the error rate is only determined afterwards, by calculating the number of misclassifications of the composite, 'jackknife' model. This is therefore different from the technique discussed by Hand (1997), where error rates are determined on each individual run, and a 'composite' error rate is determined as the estimator of the true error rate. In general, the jackknife procedure as described by Rose and Altschul and Kvamme will result in more conservative and realistic predictions than a single model built on the full data set, especially when using small samples.

Hobbs *et al.* (2002, 9-10) have used cross-validation techniques as a method to investigate the stability of their predictive model. They did not intend to produce a composite model; instead, they subdivided their data set randomly into 10 equally sized subsets, and calculated 10 different models, each time leaving out one of the subsets. These models were then compared to determine their stability, *i.e.* to see if they showed large differences. A final model could then be made using the complete data set. As a matter of fact, they were not able to carry out this cross-validation procedure as it proved too time consuming to repeat this over the 20 sub-regions of their modelling program; instead, they reverted to 'normal' split sampling. This however led to "highly unstable" models in some sub-regions, as the samples used for model building became too small. As noted above, this is not surprising, and the whole exercise clearly misses the point of using resampling methods. Instead of comparing the different models, they should have been combined in order to improve the final model.

Simple validation methods have not met with general approval in predictive modelling literature, and are not very well understood either. Ebert (2000) for example refers to 'jackknife sampling' while in fact talking about split sampling methods in general, stating that they are "a grossly inefficient way to determine if there is inhomogeneity in one's data". Gibbon (2002) notes that all testing (*i.e.* validation) methods that use the data from which the model was derived have severe drawbacks (see also Rose and Altschul 1988) – without going into detail. This did however not stop him and his colleagues from pursuing the split sampling procedure described in the previous paragraph for the MnModel.

And in fact, using split sampling for validation of predictive models is not very useful. On the one hand, split sampling will never be able to tell whether our data set is unrepresentative, as the test data is derived from the same sample as the design data. On the other hand, we should *expect* the test data set to be performing differently from the design data set. As 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 - while of course taking every precaution to prevent biases during data collection.

Resampling methods on the other hand can be valuable techniques for obtaining more realistic estimates of the accuracy and precision of a predictive model. This was already demonstrated by the jackknife sampling studies undertaken by Rose and Altschul (1988) and Kvamme (1988b; 1990). Statisticians are also quite clear that the application of resampling methods is good practice when it comes to estimating classification error. The doubts expressed on the use of internal validation methods in predictive modelling therefore have more to do with a lack of trust in the data sets used for model building, than with the applicability of validation

methods. Bootstrapping has superseded jackknife sampling as the most reliable method for error estimation. It is therefore recommended that the future analysis of classification error in predictive modelling will be done using this method, instead of jackknife sampling and cross-validation.

Unfortunately, the resampling methods described in section 8.3.2 are not very useful for validating Dutch predictive models, as they have been developed for estimating classification error, which cannot be calculated for the types of models used in the Netherlands (see section 8.2.9). It is however a logical step to use resampling techniques for the calculation of gain values as well. It can be expected that gain as obtained from the design data set will give an optimistic estimate of model performance, and therefore needs to be validated as well. In the context of this study, I have not pursued this option, and it therefore needs additional study to judge its utility.

Resampling, and especially bootstrapping[62], can also be of interest to the development of archaeological predictive models themselves, as we are usually dealing with relatively small and non-random samples. As far as is known however, it has received no attention as a modelling technique. As it is a relatively novel technique (the first textbook of bootstrap methods was published by Efron and Tibshirani in 1993) that usually requires intensive computer processing, it is not surprising that the older predictive modelling publications do not refer to it. However, there is also a complete lack of discussion of resampling methods in later publications, including some standard handbooks on sampling and statistics in archaeology (Shennan 1997; Orton 2000a). Only one reference was found in the proceedings of CAA[63] (Delicado 1999). Baxter (2003, 148-153), on the other hand, discusses bootstrapping under the heading 'computer-intensive methods', and concludes that it is generally well suited for the estimation of means and confidence intervals. He adds that caution should be applied when taking small samples from non-normal distributions and when other parameters are of interest, like the mode or median.

8.4 Statistical testing and predictive models

'from a statistical standpoint any procedure ... might appropriately be used as a basis for site-location model development. What matters, is how well the model works in application, how accurately it performs on future cases. (...) In order to determine how well a model will perform in practice ... independent testing procedures are required, and in this case methods of statistical inference must be applied.' (Kvamme 1988a)

'... perhaps it is true that all cases where the data are sufficiently ambiguous as to require a test of significance are also sufficiently ambiguous that they are properly subject to argument.' (Simon 1997)

So far, the measures and methods for model performance assessment discussed are not concerned with the significance of the outcome of the calculations. In the case where performance measures are calculated using the design data set, this is not an important issue, as there is no independent evidence to test the model against. However, when independent data becomes available, we will have to use statistical methods to decide whether we trust our model not. Section 8.4.1 will shortly discuss the utility of statistical testing methods for predictive modelling, and in section 8.4.2 some examples will be given of testing methods that can be applied to Dutch predictive models.

8.4.1 Why use statistical tests?

The testing of a site distribution against a predictive model can be used as a means to find out if there is a significant difference between the model (a statistical hypothesis) and the available data. In fact, this is what is often done as a first step in correlative predictive modelling: a statistical test is used to establish whether the

[62] Permutation resampling does not seem to be of direct relevance to predictive modeling; it assumes what is known as a 'finite universe', in which choice becomes more limited with each draw from the sample (for example the probability of obtaining a window seat in an airplane).

[63] The annual conference on Computer Applications and Quantitative Methods in Archaeology.

distribution of archaeological sites differs from a by-chance model. We are then comparing the site distribution to an uninformative statistical hypothesis[64], assuming in effect that we have no preconceptions about the possible distribution of sites. In deductive modelling or when using expert judgment models, on the other hand, we are first creating a conceptual model, and then check if the archaeological data fit the model. Testing of an existing predictive model with independently collected data does nothing different: it compares the new data set with the old model.

Such a statistical test can lead to a positive or negative result. Given a previously established level of confidence, usually the traditional 95% mark, we can state whether or not we believe that there is a difference between the model and the test data set, or to put it in more formal terms, if the null hypothesis of no difference is rejected or not. Some differences between the design and test data sets used for predictive modelling can of course be expected, for the reasons explained in section 8.3. As the model will be optimised to the design data set, the smaller this design data set is, the larger the differences may be between the model and the test data. However, a statistically significant difference between the design and test data is an indication that we are dealing with two samples with different characteristics. The only explanation for this is that we are dealing with unrepresentative samples in the design data set, the test set, or in both. Obviously, statistical testing can be a powerful tool to do precisely this: if we take every precaution to ensure that our test data set is collected according to the rules of probabilistic sampling, we will be able to tell with reasonable confidence whether or not our model was based on an *un*representative sample.

If on the other hand we find that there is no significant difference, we are ready to move on to a different aspect of statistical inference: the improvement of the model's statistical precision. The concept of estimating confidence intervals is crucial to this approach, and serves as the primary means to reduce the risks associated with statistical estimates. It is in fact the underlying principle of Bayesian statistics, where statistical models can be continuously updated with new information, each time increasing the accuracy and precision of the resulting estimates. So, statistical testing of a predictive model should consist of two phases: a hypothesis test phase, intended to identify possible biases in the original data set (at least in the case of a correlative model; with deductive/expert judgement model, the test is used to see if the model fits the data). If the model is not found at fault, a statistical inference phase follows, where the test data is integrated into the model to improve its statistical accuracy and precision.

Unfortunately, predictive models are not usually presented in the form of statistical estimates with corresponding confidence intervals. All predictive models featuring in literature only provide a relative assessment of site density. The probability of finding sites in zone A is always presented *relative* to the probability of finding them in zone B. Even logistic regression models, which do provide probabilities of site presence per grid cell, have never been used to calculate expected absolute site densities, and their residuals are never presented as a measure of the uncertainty of the model. This reluctance to present absolute site density estimates and uncertainty measures is understandable when the sample used to build the model is not based on controlled survey results but on existing site databases, like in various Dutch predictive modelling studies. Neither can it be expected from models that use expert judgment to classify zones into high, medium or low probability. However, in many American predictive modelling studies, probabilistic survey stands at the basis of model development, and absolute densities and confidence estimates could, in principle, be calculated. It is regrettable that they are not, not only from the testing perspective, but also from a perspective of cultural resource management because absolute site densities and their associated variances of course are more precise measures of the 'archaeological risk' then relative assessments. It is precisely this lack of statistical accuracy and precision in predictive models that has led to the proliferation of performance assessment measures as discussed in sections 8.2 and 8.3.

[64] In Bayesian statistics this is known as using an 'uninformative prior', and the corresponding statistical distribution is known as the uniform distribution.

8.4.2 How to test relative qualifications

Even though statistical estimates and confidence intervals are not common characteristics of Dutch predictive models, a limited form of statistical hypothesis testing is possible for relative density maps. The terms high, medium and low probability already point to a quantitative judgment: a zone of high archaeological probability is more likely to contain archaeological remains than a zone of low probability. If we want to test the accuracy of these relative qualifications of site density, we must use techniques that can deal with the *proportions* of sites that fall into each probability class. This type of testing is covered by standard statistical techniques using the properties of the binomial (in a two-class model) or multinomial distribution (in a multi-class model). The situation is somewhat complicated by the fact that we do not know what should be the actual proportion of sites in high, medium or low probability. Nevertheless, some simple tests can be used that may be helpful in deciding whether the test data set confirms or contradicts the original model.

In order to illustrate this, a simple example is presented here using the IKAW classification (Deeben *et al.* 1997; table 8.13). This classification gives the order of magnitude of the difference between the zones of high, medium and low probability: the indicative value (p_s/p_a) should be > 1.5 for high probability zones, < 0.6 for low probability zones and somewhere in between for medium probability. Let us assume that a survey has been carried out in a zone that is characterized as in table 8.13:

	area	sites found	sites predicted by IKAW rules
high probability	1 ha	3	1.5
medium probability	3 ha	2	2.9
low probability	1 ha	0	0.6

Table 8.13 Hypothetical example used for testing the IKAW-classification.

We can then calculate that, with the area proportions of the zones being 0.2:0.6:0.2, the proportions of sites in the zones should be at least equal to 0.12:0.58:0.30. So, the column 'sites predicted' gives the minimal conditions under which the model can be considered right. This makes things easier as it will allow us to establish threshold values of what the proportion of sites found in each zone should be. Taking the high probability zone, it is possible to calculate whether the 3 sites encountered might also have been found if it had in fact been a medium or low probability zone. In order to do so, we need to calculate the thresholds: in a low probability zone of the same size, we would expect at most 12% of the sites, and in a medium probability zone at most 30%. Using the binomial distribution, it can then be calculated that the probability that the high probability zone is in fact a medium probability zone is 16.3%. The probability that we are dealing with a zone of low probability (with at most 12% of the sites) is only 1.4%. Similar calculations can be carried out for the low and medium probability zones (see table 8.14).

	P(high)	P(medium)	P(low)
high probability	-	16.3%	1.4%
medium probability	1.8%	-	56.8%
low probability	16.8%	52.8%	-

Table 8.14 Probability of misclassification (P) of IKAW zones, based on the data in table 8.13.

When dealing with maps that only use a qualitative classification, our options are more limited, but we can still test whether we are dealing with a high or low probability zone, by testing the hypothesis that the high probability zone will contain more sites than expected in the case of a uniform site distribution (table 8.15). This is of course very similar to carrying out a χ^2-test against a by-chance distribution, but by using the binomial confidence intervals the probability of misclassification can be directly calculated – but only for the high and low probability. The medium probability zone already assumes a uniform distribution of sites, and in order to be 'correctly' classified, the probabilities of it being either a high or a low probability zone should be equal (50%).

	P(high)	P(low)
high probability	-	5.8%
medium probability	5.8%	94.2%
low probability	32.8%	-

Table 8.15 Probability of misclassification of qualitative probability zones, based in the data in table 8.13.

We can also try to estimate how many observations are needed in order to be sufficiently certain that we are dealing with a zone of low, medium or high probability. For this, the properties of the binomial distribution can be further exploited (see also Orton 2000a; 2000b). We can have a 95% confidence in our IKAW classification of the low probability zone (where at most 12% of our sites should be found), when at least 23.4 observations have been made, none of which falls into the low probability zone. As soon as 1 observation is made in the low probability zone, we need at least 35.6 observations in the other zones in order to accept our model at the 95% confidence level. Note that this number does not depend on the actual size of the probability zone, but on the proportion of the study region that each zone occupies.

This shows that it is very difficult to know beforehand how many observations are needed to confirm or reject the model's assumptions about site distribution when dealing with proportions. As soon as observations inside the zone of interest are found, the total number of observations needed to reject or accept the model changes. This makes it difficult to predict the actual amount of effort needed in order to reduce the uncertainty to the desired level, the more so since we do not know how long it will take to find the number of sites needed, or if we will indeed find them at all when dealing with small study regions. This is a clear argument in favour of using site/non-site models, or models based on site area estimates instead of site density (see also section 8.6.2).

A less known method of hypothesis testing is the use of simultaneous multinomial confidence intervals. Instead of testing the different classes in a model against each other, the whole model can be tested. The method is described in Petrie (1998) for an experiment using riverbed sediment size categories, but is equally applicable to any multinomial classification.

$$p_i = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

for i = 1, 2,….,k

where

a = $n + \chi^2_{\alpha/k,1}$

b = $-2n\hat{p}_i + \chi^2_{\alpha/k,1}$

c = $n\hat{p}_i$

k = the number of classes

α = the desired confidence level

$\chi^2_{\alpha/k,1}$ = the upper (1 - α/k) 100 percentage point of the chi-square distribution with 1 degree of freedom

n = the number of observations in class i; and

$\hat{p}_i$ = the observed proportion of the observations in class i.

The p-statistic can be calculated for each single class, which will result in different confidence intervals than when using the binomial intervals, the difference becoming larger when the number of classes increases. These confidence intervals can then be plotted around the observations. When, for example, a 95% confidence interval is taken, this means that the real distribution can be expected to be inside this interval 19 out of 20 times. In our hypothetical IKAW example, it turns out that the original model fits in the 71.4% confidence interval of the observations. This means that there is a 28.6% probability that the real distribution is even wider than that, implying that our small data set does not allow us to reject our original model.

8.5 Collecting data for independent testing

'It is not the modeling and it is not the sampling that makes archaeologists uncomfortable, it is the substitution for verification.' (Moon 1993)

The strongest methods of model validation and statistical testing use data collected especially for this purpose. The main source of concern for predictive modellers wanting to test their models with independent data is therefore the nature of the site sample used for testing. However, this concern is equally relevant for model building. Without representative data sets, neither model building nor testing of the model will produce valid results. This section will deal with the problems associated with obtaining representative data sets for predictive model testing (sections 8.5.1 and 8.5.2). The emphasis of this section is on retrospective model testing on the basis of so-called 'compliance' survey data, *i.e.* surveys that are carried out because there is a legal obligation to do so. In section 8.5.3 the main source of survey data in the Netherlands, contained in the ARCHIS database, will be analysed for its utility for predictive model testing. Section 8.5.4 will shortly discuss the issue of testing of the independent parameters of the model (the 'environmental data'), after which section 8.5.5 will present the conclusions on the potential of independent testing of predictive models.

8.5.1. Probabilistic sampling

All authors writing on predictive modelling agree that the collection of both design and test data sets for predictive models should ideally be done by means of probabilistic sampling, *i.e.* sampling aimed at obtaining a statistically valid sample for the estimation of *e.g.* site densities A large volume of papers and books has appeared that discuss the proper way of setting up probabilistic archaeological survey (*e.g.* Nance 1983; Altschul and Nagle 1988; Orton 2000a).

However, as Wheatley (2003) puts it: 'Data collection is precisely the activity that most model-builders are usually trying to avoid'. He may however be a bit unfair to the model builders there. After all, randomised survey has been used for predictive model building in the United States on a regular basis, and the need for independent testing is generally recognized[65]. However, it seems that survey programs for testing predictive models are not often carried out. This may be because of the associated costs, which carry the risk of becoming

[65] Even when independent testing is performed as part of model evaluation, it is not always done according to statistical principles. The tests reported by Griffin and Churchill (2000) illustrate this. Of the four surveys organized to test the Kittitas County predictive model, two failed to take into account the condition of representativity, and over-sampled the high probability area. Furthermore, even though the two reliable survey tests indicated that the model was not accurate, it was not adapted.

structural expenses, whereas the intention of a predictive model is to reduce the amount of money spent on survey. In any case, this is not a very strong argument: if money is invested in collecting data to build the model, then certainly some money for testing can be made available. A second possible cause for the lack of testing programs may be that predictive modellers have not been able to communicate exactly how much data is needed for testing a model. In order know how many observations are needed for our test, we should be able to specify both the actual statistical precision of the model, as well as the desired precision. Furthermore, as already shown in section 8.4.2, even when we know how many site observations we need, we cannot know beforehand how many hectares need to be surveyed, as this depends on the actual site densities involved – the very parameter that we are trying to estimate.

Because of the lack of probabilistic testing programs, it is almost inevitable that the so-called 'compliance' surveys form our main source of independent data. In general, these are not carried out with a probabilistic aim in mind. Their main objective is the discovery of all, or a predetermined proportion, of the archaeological sites in an area. This is known as 'purposive' sampling, and some authors reserve the term prospection for it. Purposive survey has been less debated in academic literature, and only Banning (2002) makes a clear distinction between the two.

An important obstacle to the use of compliance survey data for testing purposes is the difficulty of collecting data from many small survey projects. 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 compliance 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 and Churchill (2000) for an example from practice; Wheatley (2003) for a critique of this approach; and Verhagen (2005) for some examples of institutionalised bad habits). Other sources of bias originate from survey practice (see section 8.5.2 for more details). 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. As Altschul and Nagle (1988) already remarked: "The failure to utilize inventory survey results is not only an unfortunate decision but also in the long run an extremely expensive one". However, the only way to use these data sets for predictive model testing is to analyse the different sources of bias and, if possible, correct for them.
From a classical statistical standpoint, 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;
- and survey methods should be chosen such that bias in site recording is avoided.

The standard procedures to calculate appropriate sample sizes can be found in any statistical handbook (*e.g.* Shennan 1997; Orton 2000a), and are based on the assumption that samples consist of two classes, like site presence-absence counts per area unit. In Dutch predictive modelling however we are dealing with point observations of sites: samples with only one class. Furthermore, we do not know the proportion of the area sampled, which makes it impossible to specify statistical estimates and corresponding confidence limits of site density. In section 8.4.2 it was shown that we can to some extent test the classification of Dutch models like the IKAW in low, medium and high probability using single class samples. It is however very difficult with these models to determine in advance the number of observations needed, as it also depends on the observations done in the other probability zones. Furthermore, we cannot predict the size of the area that should be sampled in order to obtain the required sample size, as we do not know the real site density in the survey area.

An additional problem for predictive model testing is the low number of sites included in the low probability zones. A reliable estimate of site densities in the low probability zone requires more data collection than in the high probability areas. This is because the estimates of low numbers can be substantially altered by the discovery of very few sites; the estimates are less stable.

This again points to the importance of making models that do specify statistical estimates of site density and confidence limits. Probabilistic sampling can evidently be used to provide these estimates, especially since the size of the survey quadrats is usually determined in such a way, that site density estimates per area unit can easily be obtained, *e.g.* per hectare or square km. Compliance survey on the other hand usually is carried out in contiguous parcels with unequal sizes, but when the positions of individual observations are known, a raster GIS can be used to create equal sized sampling units. Note however that the size of the sampling unit has a strong effect on the calculation of the confidence limits. The creation of very small sampling units implies that the sample may become artificially large, which is paraphrased by Hole (1980, 226): "by planning infinitely small sample units, one could know everything about an area by looking at nothing". A possible solution is to stop creating artificial sampling units, and instead use resampling techniques to calculate density estimates and confidence limits from the site observations in the total sampled area.

8.5.2 Survey bias and how to control for it

Unfortunately for predictive modellers, there are other sampling issues that must be taken into account as well, and especially the influence of survey bias. Even more regretfully, methods and procedures for controlling and correcting survey bias have not featured prominently in or outside predictive modelling literature, although *e.g.* Shennan (1985) and Verhoeven (1991) tried to identify and quantify sources of bias in field survey data with statistical techniques (see also van Leusen 2002 and Attema *et al*. 2002). The main sources of bias identified are:

- the presence of vegetation, which obscures surface sites;
- sediment accumulation, which obscures sub-surface 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 (2000a) identifies imperfect detectability as the main source of non-sampling error in archaeological survey (the subsidiary source being non-response). Correcting site density estimates for imperfect detectability is relatively easy, using the following equations (Orton 2000a, 214-215):

$$\hat{\tau} = \frac{\tau_0}{g}$$

$$v(\hat{\tau}) = \frac{v_0}{g} + \frac{\hat{\tau}(1-g)}{g}$$

where

$\hat{\tau}$ = the corrected estimate
τ_0 = the original estimate
$v(\hat{\tau})$ = the corrected variance
v_0 = the original variance
g = the detection probability.

These equations will result in higher estimates of site density, with a larger variance.

The task of bias correction then becomes a question of estimating the detection probability of a particular survey. Obviously, this would be easiest when survey results were based on the same methods. 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, given in Orton (2000a, 211-212). 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; 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, and consequently the survey methods could not be evaluated for their effectiveness. However, in the context of predictive model testing, a way out 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 these 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.

8.5.3 Using the ARCHIS database for predictive model testing

The most accessible archaeological data set available in the Netherlands is the ARCHIS database. The data in ARCHIS is structured in the following way:

- any form of archaeological research has to be registered before it starts; however, this obligation has only started early 2004; in addition, the curators of ARCHIS are actively filling the database with the backlog of archaeological research carried out before 2004; at the moment of writing (4 March 2005), 9,043 research projects have been registered.
- the completion of archaeological research must be registered as well; at the moment, 5,104 completed research projects have been registered.
- any archaeological observation made must be registered; in practice, this has not been enforced, but since the start of the ARCHIS database in 1991, a total of 65,944 observations have been entered, including those from many paper archives.
- the number of archaeological complexes ("sites") however, is only 17,066
- in addition, the database contains 12,941 archaeological monuments.

The archaeological observations are coming either from archaeological fieldwork, from paper archives, or from non-archaeological fieldwork (table 8.16).

archaeological fieldwork	36,481	55.3%
desk-top study and archival research	4,051	6.1%
non-archaeological fieldwork, including metal detecting	18,413	27.9%
not specified	6,999	10.6%

Table 8.16 Breakdown of archaeological observations in ARCHIS according to discovery.

The observations made during archaeological fieldwork can be subdivided into several categories (table 8.17. Most observations made by archaeologists are coming from field walking and excavation. If we look at the number of registered research projects however, the picture is quite different (table 8.18).

core sampling	2,526	6.92%
field walking	23,788	65.21%
watching briefs	257	0.70%
diving	2	0.01%
geophysical survey	48	0.13%
archaeological inspection	2,058	5.64%
not specified	1,085	2.97%
underwater survey	23	0.06%
test pits/trial trenches	317	0.87%
excavation	6,377	17.48%

Table 8.17 Breakdown of archaeological observations in ARCHIS found by archaeological fieldwork, according to research type.

These data show that core sampling is taking up the vast majority of archaeological fieldwork nowadays, with test pitting/trial trenching, watching briefs and excavation gaining in popularity (the 'completed' column should be read as the research preferences over the past 10 years or so, the 'registered' column as the current preferences). Incidentally, a quite staggering number of 6,432 unspecified research projects (71.1% of the total) has been registered. We can only assume that these will be attached to one of the fieldwork categories once the fieldwork is finished. As the ARCHIS curators are at the moment still working at reducing the backlog, and the amount of research done nowadays is enormous, the figures cited in table 8.18 may change considerably in the near future.

Unfortunately, it is impossible to obtain from the ARCHIS database the data needed for the development of an acceptable predictive modelling test data set. Registration of the fieldwork projects is 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. Furthermore, a major problem with the delimitation of study areas becomes apparent in the municipality of Het Bildt (province of Friesland), which contains 26 database entries, covering

the entire municipality, and the neighbouring municipality of Ferwerderadeel, which has another 34. These 60 projects together are taking 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 IKAW-zones or archaeo-regions. The existing data points to a preference of all types of archaeological fieldwork towards the high potential zones, with the exception of geophysical survey. This preference becomes more marked when moving from inventory survey (core sampling, field walking) to evaluation and excavation. Watching briefs are the most representative form of archaeological research, and these results conform expectation. However, given the misgivings regarding the quality of the data a quantitative analysis of the research database has not been pursued.

projects	registered		completed	
core sampling	1,678	66.51%	3,885	77.50%
field walking	59	2.34%	115	2.29%
watching briefs	170	6.74%	158	3.15%
diving	0	0.00%	0	0.00%
geophysical survey	17	0.67%	215	4.29%
archaeological inspection	4	0.16%	74	1.48%
not specified	11	0.44%	22	0.44%
underwater survey	4	0.16%	3	0.06%
test pits/trial trenches	343	13.59%	375	7.48%
excavation	237	9.39%	166	3.31%
total	2523		5013	

Table 8.18 Breakdown of registered research projects, according to research type.

The ARCHIS research projects database was never intended for use as a test set for predictive modelling, and cannot be used directly for this purpose. We are forced to return to the original research project documentation to find out which areas have actually been surveyed, and which methods have been applied. This task, which has not been possible within the constraints of this study, should be a priority for future improvement of the IKAW.

From a statistical point of view, the representativity of the data is of course important, but equally important is the total number of observations obtained from the various forms of survey, because this determines whether the fieldwork data can actually be used for testing purposes. Here the picture (see table 8.19) is not very promising either. Of the 4,155 observations registered since 1997 (the publication date of the first version of the IKAW), only 1,589 can be linked in ARCHIS to a registered research project (*i.e.* they are found within an investigated zone, and the observations have been classified into one of the archaeological fieldwork categories). Given the fact that the original model was developed for 13 separate 'archaeo-regions', on average just over 120 observations per region will have to be analysed by survey type in order to remove research biases. Serious

doubts should therefore be expressed concerning the current value of these observations for rigorous predictive model testing.

observations	low probability	medium probability	high probability	unspecified	TOTAL
core sampling	114	157	225	65	561
field walking	100	140	205	13	458
Watching briefs	14	29	34	8	85
geophysical survey	-	5	-	-	5
archaeological inspection	3	6	3	1	13
not specified	8	8	15	4	35
test pits/trial trenches	39	44	99	23	227
excavation	36	51	103	37	205
total	314	440	684	151	1,589

Table 8.19 Number of archaeological observations made in research projects since 1997, subdivided by research type and archaeological probability zone on the IKAW.

However, even from these unsatisfactory data, a surprising pattern is found in the distribution of observations registered from test pits/trial trenches and excavations: they yield more sites in the low potential areas of the IKAW than expected[66]. Even though excavation or trial trenching in low potential areas will only be done when the presence of a site is known or suspected, this is not any different for the medium and high potential zones, and should therefore in theory not lead to higher discovery rates. For the moment however, the data does not suggest an explanation for this observation.

8.5.4 Testing the environmental data

In practice, Dutch predictive models are not tested in a quantitative sense. 'Testing' a predictive model is usually understood to mean verifying and refining the environmental base data underlying the model, like soil maps or palaeogeographic information. Changes in these data do find their way into the models. The most important reason for this is that the relevant environmental information is often much easier to detect than the archaeological sites themselves. With core sampling, for example, it may be hard to find certain types of sites, but finding the extent of a particular geological unit that was used to construct the predictive model is relatively easy, and may serve to make the model more precise. In fact, it is a question of improving the scale of the mapping for the predictive model, as well as reducing errors in the base data. On 1:50,000 scale maps for example, all kinds of generalizations have been performed, and the base data used may exist of fairly widely spaced observations. When prospecting for archaeological sites, the level of detail is much finer than with standard geological or pedological mapping, and archaeological prospection therefore contributes to a better knowledge of the (palaeo-)environment as well.

Getting these new environmental data back into the predictive map may pose some problems. A change in the patterning of the parameters used for the model in fact implies that the whole model should be re-run.

[66] Excavations: 18.9% instead of < 7.3%; trial trenches: 21.4% instead of < 9.0%.

When using an expert judgement model, this is relatively simple: a new map can be drawn quickly, using the new information obtained, and depending on the nature of the changes, the model will become more or less precise and accurate. However, as soon as we want to use the new information in a quantitative model, the whole modelling procedure should be rerun to see if the site patterning changes. An additional problem is found in the fact that this type of testing is seldom done for the whole area covered by the model. There are rare instances where the independent parameters of predictive models are completely revised, *e.g.* when better soil maps have become available (Heunks 2001), or a new detailed elevation model has become available (*e.g.* van Zijverden and Laan 2005). In most cases however, we are dealing with very limited testing, that will be difficult to feed back into the model, as the result will be a model based on data with differing resolutions.

8.5.5 Conclusions

The perfect data set for predictive model testing is one that is representative of the area, and does not have the problem of imperfect detectability. From this it follows that data obtained from watching briefs is best suited for testing, as it permits for the observation of all sites present as well as the areas of non-site; at the same time it is not limited to the zones of high probability. Unfortunately, the registered number of discovered sites by means of watching brief operations in the Netherlands is very low, and it also seems that watching briefs are now increasingly used as a substitute for trial trenches or even excavation (they are now formally known as 'excavations with restrictions'). Originally, a watching brief was carried out as a final check on the presence of previously unnoticed or 'unimportant' archaeology. Nowadays, they also have become obligatory procedures in cases where the presence of an important site is known or suspected, but the damaging effect of the development is considered too minimal to justify a full-scale excavation. These types of watching briefs are not particularly useful for predictive model testing, as they will not be representative of the test area.

Given the low number of watching briefs available, retrospective predictive model testing will (also) have to rely on other data. The most abundant data set available nowadays is core sampling survey data, and it is therefore logical to concentrate our efforts on this data source. Furthermore, it can be expected that, together with field walking, it will suffer less from the effect of unrepresentative sampling than trial trenching and excavation. Even though the effect of imperfect detectability will to a certain extent distort the estimation of the number of sites discovered, these effects can be analysed and corrected for if the core sampling strategy used (depth, coring equipment used and spatial layout of the survey) is registered. Obviously, this still is a lot of work, the more so because the registered research projects database in ARCHIS cannot be relied upon to contain these data for each registered project.

For future testing, it is imperative that the registration of the survey strategy followed, in terms of sampling unit size, layout and surveyed area, is done correctly, and preferably stored in a central database. For pro-active testing, it is essential that it is done according to the principles of probabilistic sampling, and that the size of the data set to be collected is based on the desired precision of the estimates needed.

8.6 The test ground revisited

In the preceding sections, a number of issues have been raised concerning the best ways to test predictive models. It will have become clear that the applicability of testing methods highly depends on the type of model under consideration. Section 8.6.1 will therefore summarize the appropriate testing methods for the main types of predictive models that are currently in use. Section 8.6.2 will then continue with what I consider to be an alternative method of predictive modelling, *i.e.* modelling using area estimates instead of site (and non-site) counts. I will argue that this type of modelling is more useful for archaeological risk assessment than traditional predictive modelling approaches.

8.6.1 Model types and appropriate testing methods

In practice, we can distinguish five major types of predictive modelling procedures that are currently used. The following scheme summarizes their main characteristics:

expert judgment / intuitive models (example: Heunks *et al.* 2003)

- single-variable; multiple variables are combined intuitively into new, composite categories
- changes in the independent parameters can be accommodated easily by redrawing the base map
- classification of categories into high-medium-low
- no quantitative estimates
- no confidence limits
- gain and gain-like measures can be used to assess model performance, using a test data set
- the precision of the model can be increased by reducing the high potential area
- statistical hypothesis testing is limited to deciding whether the model correctly predicts if unit A has a higher/lower site density than unit B

deductive / expert judgment multi-criteria analysis models (example: Dalla Bona 1994)

- multivariate; combinations of variables by means of Boolean overlay
- changes in the independent parameters can be accommodated relatively easily, but imply running a new model
- classification of categories in 'scores', that can be translated to high-medium-low
- 'scores' are not estimates
- no confidence limits
- gain and gain-like measures can be used to assess model performance, using a test data set
- the precision of the model can be increased, by manipulating the scores
- statistical hypothesis testing is limited to deciding whether the model correctly predicts if unit A has a higher/lower site density than unit B

correlative / inductive site density transfer models (example: Deeben *et al.* 1997)

- single-variable or multivariate; combinations of variables by means of Boolean overlay
- changes in the independent parameters can be accommodated relatively easily, but imply running a new model
- classification in categories of relative site densities (using indicative value or other measures)
- quantification in relative terms
- no confidence limits
- the use of performance measures as well as performance optimisation is inherent to the modelling procedure
- statistical hypothesis testing needs an independent test data set, that can be used to
  - decide whether the model correctly predicts the relative site densities in zone A compared to zone B
  - decide whether the total model differs from the test data set

correlative / inductive regression models (example: Hobbs *et al.* 2002)

- multivariate, using logistic regression
- changes in the independent parameters cannot be accommodated without doing a new regression analysis
- probability values of site and non-site
- classification into site-likely and a site-unlikely zone (two-class)
- no confidence limits
- performance measures can be used, but should preferably be based on a test data set; apart from gain-like measures, classification error can be used as a measure of model performance
- performance optimisation can be used after the regression analysis
- statistical hypothesis testing needs an independent test data set, that can be used to
  - decide whether the total model differs from the test data set

Bayesian models (example: Verhagen 2006)

- single-variable or multivariate, based on a statistical hypothesis (a priori distribution); this hypothesis can be based on expert judgement, or on an existing data set
- changes in the independent parameters cannot be accommodated without running a new model
- estimation of site densities, either absolute or in proportions, and corresponding confidence intervals, that can be reclassified into 'crisp' categories
- gain and gain-like measures can be used to assess model performance after reclassification
- performance optimisation can be used after the modelling
- statistical hypothesis testing needs an independent test data set, that can be used to
  - decide whether the total model differs from the test data set
  - integrate the new data into the model

Note that the first four are categories of models that have been used extensively for predictive modelling, whereas models of the Bayesian type are far from generally applied. Nevertheless, these are the only type of model that will automatically result in statistical estimates with corresponding confidence intervals, and are mathematically best suited for statistical hypothesis testing purposes as they include a mechanism to deal directly with the results of the test. In fact, it is a question of feeding the new data into the model, and it will be updated automatically.

It should also be pointed out that the fact that logistic regression models are presented here as models that do not specify confidence intervals should not be taken to mean that these cannot be calculated. In fact, the calculation of what is generally known as the 'error term' of the regression equation is common statistical practice, and it is a mystery why it is not customarily included as a measure for model performance in published archaeological predictive models. A significant advantage of regression techniques is that they can include measures of spatial autocorrelation as well, and methods to do so have already been developed in other fields.

Concerning other types of models that have recently attracted some interest, it can be remarked that both land evaluation-based models (Kamermans 2000; 2003) and causality-based cognitive modelling (Whitley 2003; 2005a) are from the technical point of view comparable to multi-criteria analysis models. Dempster-Shafer theory, used for predictive modelling purposes by Ejstrud (2003; 2005), is at the conceptual level connected to Bayesian modelling. However, it is still very much debated in statistical literature, as it is not considered a true probabilistic technique (see Smets 1994 and Howson and Urbach 1993, 423-430). Dempster-Shafer models result in probability values just like regression models, but they also provide a measure of uncertainty known as the 'belief interval'. From the example given by Ejstrud (2003), it is not immediately clear what testing methods are appropriate for these models. Gain calculations can be done on the models, and given the parallel with Bayesian methods it can be assumed that Dempster-Shafer models can easily be updated with new information in order to reduce uncertainty.

It thus turns out that performance assessment by means of gain and gain-like measures is the only kind of test currently available to all types of models. This also means that these measures are the only ones that can be use to compare different modelling techniques, as has been done by Ejstrud (2003). Obviously, the kappa coefficient (see section 8.2.6) can be used for comparison purposes as well, but it will only point to a difference between models, and will not indicate if model A is better then model B. A major disadvantage of using gain and gain-like measures as the sole indicator of model quality is the fact that they cannot be used to predict the model's performance for future cases. For this, we need methods of statistical inference, and models that provide actual statistical estimates with confidence intervals. This implies that for each model, correlative or not, a representative data set should be available from which to make these estimates.

The development of resampling techniques allows us to obtain statistical estimates and confidence intervals per probability zone from a test data set for all model types. As such, resampling can provide a valuable contribution to model performance assessment. Resampling may equally well be used to develop correlative predictive models that provide estimates and confidence intervals, and at the same time do not need the complex

statistical hypotheses necessary for the proper use of Bayesian models. In fact, the great advantage of resampling techniques is that they do not presuppose a specific statistical distribution at all. However, resampling needs further investigation to judge its ability to be applied to multi-variate models, to see if it can be combined with expert judgement and deductive modelling, and how it can be used to make comparisons between models.

8.6.2 Towards an alternative form of predictive mapping: risk assessment and the use of area estimates

'*It is impossible to say anything about the number of archaeological phenomena that can be expected other than in terms of "relatively many" or "relatively few*' (Lauwerier and Lotte 2002, referring to the IKAW)

As a final consideration, I will devote some attention to the issue of predictive modelling and risk assessment. A predictive model is a 'decision rule': the only reason for making a predictive model in archaeological heritage management is that it will allow us to make choices on the course of action to be taken. In order to do so, the model must distinguish between zones that are more or less important, and each individual zone (implicitly) carries with it a different decision. We are dealing with risk assessment here, and quantitative methods are obviously well suited to contribute to this analysis. From this perspective, it does not really matter what we distinguish on a predictive map, but only how effective it is in supporting the decisions made.

In current practice, predictive models are used to make comparisons between development plans, and to decide whether or not an area should have some form of legal protection, in the form of an obligation to do survey. In itself, this is not bad practice from the point of view of risk management. If the model is correct, the high potential zones will contain more sites, and will therefore have a higher 'production' of archaeology (see also Banning 2002). The return, or 'archaeological profit', of prospection in high potential areas will therefore be higher than in low potential areas, given the same prospection intensity.

Archaeologists however are generally reluctant to accept the fact that interesting archaeological phenomena may escape detection as a consequence of a risk assessment carried out on the basis of predictive maps. They would be much happier with predictive maps that only depict the zones of no probability, in other words, models that do not exhibit gross error. These no-probability zones would on the one hand constitute the zones where habitation has not been possible, and on the other hand the zones that are known with certainty to have been disturbed. All the other zones then need survey in order to establish the presence or absence of sites. In the current political circumstances, this is not an acceptable policy, as it will lay an archaeological claim on almost all development plans. We will therefore have to accept that risk assessment is carried out based on predictive maps, and will lead to the designation of zones where archaeological survey is not obligatory, even though we know that archaeological sites may be present. The only thing we can try to do is reduce this risk to the minimum.

It is therefore the definition and calculation of archaeological risk that should bother us, and that should be at the heart of predictive models. However, none of the currently used models comes close to defining the archaeological risk in such a way that it can be used for effective decision-making. Even before making a predictive map we should already think about the acceptable 'archaeological loss'. Much of this is related to the issues discussed in this chapter. We can establish quality criteria for predictive maps, stating that *e.g.* 85% of all known sites should be found in the high probability zone, thereby accepting that in future cases about 15% of the archaeological sites may be lost without investigation. We can also establish as a quality criterion that this 85% should not only be true for the currently known site database, but also for future cases. In other words, we need to make statistical estimates of the total number of sites, based on representative survey data sets. When the statistical estimates show that there is a large amount of uncertainty, then we can develop survey programs to reduce this uncertainty. This is however only part of the equation. Some sites are considered more valuable then others, some sites are more expensive to excavate than others; basically, we should establish priorities of what we want to protect, either for scientific, aesthetic or financial reasons. It is only after establishing these

priorities that a predictive model can be used for risk assessment. This also implies that different priorities will lead to different outcomes of the assessments, and that these may become quite complex.
One consequence of this development is that we should start thinking about different ways to produce predictive maps, by incorporating the priorities for archaeological risk assessment into the model. One option that needs to be explored is the possibility of making predictive maps that are not based on site density, but on the area taken up by significant archaeological remains (Orton 2000b). Archaeological sites can have astonishingly large variations in size and content. Lumping them all together in one model is not justified from the perspective of archaeological science, and also offers a problem for archaeological heritage management. Clearly, different types of sites (or off-site phenomena) ask for different strategies of prospection, evaluation and eventually excavation. Different site types therefore also have different associated costs – some sites are simply more expensive than others, regardless of their scientific and/or aesthetic value. There is very little published information available on the variables that determine the 'price' of an archaeological site, and some would probably consider it unethical to treat them this way. But the simple fact is that this issue will be of great interest to the developers who have to pay for archaeological research. Clearly, the current state of predictive modelling does not allow us to put an 'archaeological price tag' to a development plan. It is very hard to determine for each and every site type how likely they are to be found in a specific zone, but at least one characteristic is worth considering for inclusion in predictive models, and that is the size of the site. Not only is it one of the main determining factors for establishing the cost of archaeological research, it is also relatively easy to use for making statistical estimates. Instead of predicting the expected number of sites, a predictive model would then have to predict the expected area taken up by archaeology.

However, when using area estimates of significant archaeological remains instead of site counts, we are confronted with some statistical subtleties. Standard statistical techniques deal with populations that can be counted, *i.e.* they can be subdivided into clearly distinguishable objects (the traditional red and white balls in the urn). From samples of these objects, properties can be measured and statistical estimates can be made. As statistical theory allows for the calculation of proportions as well as of totals, one could use *e.g.* a sample of trial trenches to calculate the expected total area and corresponding confidence intervals of site, off-site or non-site. However, trenches are often of unequal size, and unlike with the parcel survey problem described in section 8.5, we do not have predefined units from which to sample[67], so effectively we do not know the size of the sampling unit population. A possible approach to tackle this problem is to consider the smallest possible unit (the smallest trench) as the principal unit of investigation, and calculate the total number of sampling units from it. The size of this smallest unit will have a strong effect on the resulting estimates and confidence intervals, as this depends on the number of samples rather than the total area covered by trenches (see section 8.5; and Orton 2000a, 25).

The solution to the problem is offered by the calculation of ratio estimates (Cochran 1963, 29-33). The ratio we are interested in is the proportion of 'site' in the excavated area. The estimation of this ratio is of course very simple, as

$$r = \frac{\sum_{1}^{n} y_i}{\sum_{1}^{n} x_i}$$

where

r = the ratio estimate obtained from a sample of size n
y = in this case, the area of significant archaeological remains
x = in this case, the researched area per project

[67] This is not to say that this cannot be done – in fact, for reliable estimates we had better develop trenching campaigns that do use equal-size units; it is just not common practice.

In order to calculate the standard deviation of the estimate, the following equation can be used:

$$s_r = \frac{\sqrt{1-f}}{\sqrt{n \cdot \bar{x}}} \cdot \sqrt{\frac{\sum y_i^2 - 2 \cdot r \cdot \sum y_i x_i + r^2 \cdot \sum x_i^2}{n-1}}$$

where

f = the finite population correction factor

$\bar{x}$ = in this case, the mean researched area

These equations can be used for estimating area totals from uneven sized trenches, or from different survey projects. For illustration purposes, a small test was carried out with these equations, using trenching data from the Midden-Delfland project[68]. In one of the study regions (Module 7 Polder Noord-Kethel), a total of 206 trenches were dug. Even though the excavation data did not permit to clearly delimit 'sites', it is well suited for determining the total area of archaeological features. The total area excavated is 5950 m^2, equating to a proportion of 0.54% of the study region. This latter figure can be used as the finite population correction. Within the trenches, 115.5 m^2 of (Medieval) features were recognized and registered, equating to 1.9% of the excavated area. The estimate of the proportion of features dated from the Medieval Period in the study region is therefore equal to 1.9%, and the calculated standard deviation is 1.3%. A 95% confidence interval can then easily be calculated by multiplying the standard deviation with the corresponding z-value of 1.96, resulting in 2.5%. This is a relatively narrow estimate, as a consequence of the large number of trenches, even though the total area coverage is far below the generally recommended coverage of 5% for site evaluation.

In a similar way, estimates can be produced from compliance surveys: the area that is eventually selected for protection or excavation then is the area that we are interested in. A first impression how large that area is can be obtained from the data that has been collected by Past2Present/Archeologic[69], in order to quantify the archaeological risk of development plans. From their data, on a total of 23 projects, it is estimated that 23.9% of the area of a development plan where compliance surveys have been carried out will be selected for protection or excavation. The 95% confidence interval of this estimate is then 13.3%.

8.7 Conclusions and recommendations

8.7.1 Conclusions

The baseline report of the project 'Strategic research into, and development of best practice for, predictive modelling on behalf of Dutch cultural resource management' (van Leusen *et al.* 2005) identified four main issues to be investigated under the theme 'Testing'. These were:

- designing test protocols in order to assess the quality of predictions/zonations
- studying the potential of 'retrospective testing' through the use of archaeological data generated by recent large infrastructural works
- studying the feasibility of proactive testing through a programme of coring, test pitting and trial trenching
- studying the potential of available statistical testing techniques such as 'jackknife' and other methods

The results of the current study suggest that:

- the design of test protocols depends on the type of model used, and the quality criteria established. Given the fact that quantitative quality criteria for Dutch predictive maps are absent at the moment,

[68] Kindly put at my disposal by Heleen van Londen (AAC Projectenbureau, University of Amsterdam).

[69] Used here with their permission.

only general guidelines can be supplied in this respect:
    - predictive model performance should be calculated using an external data set rather than the design data set;
    - any external data set used for predictive model testing should be collected according to the principles of probabilistic sampling; and
    - the most powerful testing methods imply the use of statistical techniques that can specify the uncertainty of site density estimates.
- retrospective testing of the currently used predictive models in the Netherlands is hampered by the lack of reliable and easily accessible documentation of the available archaeological data sets. Collecting and analysing the available data sets is possible, but will entail a major effort.
- for proactive testing, the same holds true. The objective of new data collection should be to obtain a representative test data set of sufficient size. This implies that surveys should also be carried out in areas of low probability. The appropriate size of the data set should be calculated from the confidence intervals of site density estimates, and depend on the desired precision of the estimates. So, without analysing and using the data sets available for retrospective testing, we will not be able to say where we need to collect new data, and how much. However, if we start with small proactive testing programs, it should be possible to slowly improve the models' quality by integrating new data when it becomes available.
- resampling offers the potential of obtaining statistical estimates and confidence intervals, even from relatively small data sets, and with any type of predictive model. It therefore is a promising technique that needs further development for predictive modelling purposes.

A number of additional conclusions can be drawn on more technical issues:

performance assessment

accuracy and precision:
- a good predictive model should be both accurate and precise, *i.e.* the high probability zones should capture as many archaeological sites as possible in as small an area as possible.
- the accuracy and precision of any predictive model can be calculated using a number of gain measures that are applicable to all types of predictive models. However, none of these measures fully solves the problem of deciding whether model A is really performing better than model B, as this also depends on the quality criteria imposed on the model (*i.e.* whether accuracy is more important than precision, and how much more important).

classification error:
- the calculation of classification error, while a statistically more sophisticated measure of model quality, is only possible when using a binary, site/non-site model. Performance measures based on misclassification rates can therefore not be applied to currently available Dutch predictive models, which do not use a site/non-site approach.

model optimisation:
- various model performance optimisation methods have been developed for quantitative predictive models, and allow for a trade-off between accuracy and precision. The use of these methods implies that the maximum performance possible of the model can be obtained with the available data set.
- with qualitative models, only the precision can be manipulated by changing the weights of the probability zones.
- in the case of site/non-site models, the 'intersection method' can be used for optimisation, allowing for

a trade-off between gross and wasteful error.
- with site density transfer models, gain development graphs are practical tools to distinguish between zones of low, medium and high probability.

quality norms:
- model performance criteria are in most cases not defined, making it impossible to decide whether the model is accurate and precise enough.
- however, accuracy is, from the point of view of protection of the archaeological heritage, a more important criterion of model performance than precision. Low probability zones are preferred by developers because of the low archaeological risk in these zones, and will usually be surveyed less intensively. An inaccurate model will therefore increase the risk that archaeological sites are destroyed unnoticed.

model validation:
- validation implies the calculation of performance measures using either new data (double validation) or parts of the design data set (simple validation).
- validation will give a more realistic indication of model performance than is obtained by calculating performance statistics from the design data set itself.
- split sampling keeps parts of the design data set behind, to obtain a semi-independent check of model performance.
- resampling techniques (including jackknife sampling and bootstrap sampling) re-use parts of the design data set to obtain model performance measures, and are closely related to Monte Carlo simulation techniques.
- when using validation, bootstrap sampling is the most sophisticated technique available, and should therefore be used in favour of other techniques.
- all validation techniques discussed are primarily intended to obtain more realistic estimates of the classification error. These techniques can therefore not be used directly with current Dutch predictive models. However, it is possible to use them for other purposes as well, like the calculation of gain.
- apart from its application as an error estimation technique, resampling is a new and rapidly growing branch of statistics that allows for statistical inference in situations where sampling conditions are far from ideal. However, archaeological applications are virtually absent up to now.

statistical testing:
- in order to apply statistical testing to predictive models, they should provide estimates of site densities or proportions, and the corresponding confidence intervals.
- statistical testing of correlative predictive models should consist of two phases: a hypothesis test phase, intended to identify possible biases in the original data set. If the model is not found at fault, a statistical inference phase follows, where the test data is integrated in the model to improve its statistical accuracy and precision.
- two types of model are suited for this type of statistical inference: Bayesian models, and models based on resampling. Bayesian models have not yet left the stage of pilot studies, and resampling has never even been considered as a tool for predictive model building.
- in all fairness, it should be added that logistic regression models are also open to statistical testing and improvement, but they are not normally presented with confidence intervals or error terms that may be reduced by using a test data set.
- most currently available predictive models however only provide relative site density estimates. Suitable statistical testing methods for this type of models are extremely limited. This is the primary reason why performance measures are used to assess model quality instead of statistical tests.

independent testing

data set requirements:

- in order to perform any kind of test, the test data set should be collected according to the rules of probabilistic sampling. This does not mean random, but representative sampling.
- the use of site counts as the basis for predictive models makes it difficult to decide how much data collection is needed for testing, as we do not know how long it will take to find the number of sites needed to reduce uncertainty to an acceptable level. This is a clear argument in favour of using site/non-site models, or models based on area estimates instead of site density only.
- a reliable estimate of proportions in low probability zones requires more data collection than in high probability areas. This is because the estimates of low numbers can be substantially altered by the discovery of very few sites; the estimates are less stable.

retrospective testing:

- for retrospective testing, representative data are often not available. The use of non-probabilistic sampling data for retrospective testing purposes is possible, but needs careful data analysis and correction of biases.
- the least biased data source available for retrospective testing is the watching brief. Given the relatively low number of watching briefs carried out, it is inevitable that other survey data will need to be analysed as well. Core sampling data is the most abundant data source available.
- the current data contained in ARCHIS is not well suited for predictive model testing. Errors in data entry as well as insufficient registration of the potential sources of bias of the research carried out make it impossible to carry out a reliable test of, for example, the IKAW. The database could however be used to find out which projects have been carried out in a specific area, so the relevant documents can be collected. This data has to come from many different sources, and will not all be available in digital form.

8.7.2 Recommendations

From the current study, the following recommendations result:

- the Dutch archaeological community should define clear quantitative quality criteria for predictive models. The accuracy of predictive models, measured as percent correct prediction of an independent and representative archaeological data set, should be the first concern of heritage managers.
- performance measures of archaeological predictive models should always be calculated using a representative test data set, and should be calculated for correlative models as well as for expert judgement models.
- validation by means of resampling, specifically bootstrap sampling, is good statistical practice for the calculation of performance measures, and should become part of model quality assessment procedures. However, the potential of statistical resampling methods still needs further investigation, for archaeological predictive model building as well as for testing purposes.
- ideally, both correlative and expert judgement/deductive predictive models should be able to provide statistical estimates and confidence limits of site density or site area for the probability zones distinguished. This allows for the establishment of the desired confidence limits, which can be used as a criterion for the amount of future data collection needed.
- once quality criteria are available, formal testing protocols can be developed to perform quality control of the models, and research programs can be developed to reduce the uncertainties in the current models. Preferably, these research programs should be embedded in the normal procedures for archaeological heritage management, *e.g.* by specifying the amount of testing to be done in project briefs for compliance surveys. In fact, it implies that probabilistic sampling should be done *together*

with purposive sampling.
- coupled to this, the quality of the archaeological data sets used for building correlative models should always be analysed. They should be representative of the area modelled, and survey biases should be detected and corrected for. Correlative models based on biased, unrepresentative data should not be used.
- it is open to debate whether models based on statistical estimates and confidence limits will lead to better predictions, and therefore to better decision making in archaeological heritage management. However, as we do not have any such models available right now, this cannot be judged. It is therefore recommended to start a pilot study, using different modelling techniques and a test data set, to compare the performance of these techniques with traditional models.
- from the perspective of risk management, models that predict the area taken up by significant archaeological remains are more useful than site density estimates. These models can in fact be made, but require a substantial amount of data collection and analysis. Again this needs a pilot study, in order to judge the feasibility of such an approach.
- the main archive of archaeological data in the Netherlands, the ARCHIS database, is not well suited for predictive model testing. Especially the research database should be thoroughly analysed and corrected in order to obtain representative site samples.
- future data entry procedures in ARCHIS should take into account the level of information needed for predictive model testing. This primarily includes the correct delimitation of the zones investigated, and the registration of factors influencing detection probability: the fieldwork methods used, the size and configuration of sampling units, the depth of investigation, and factors that cannot be manipulated, like vegetation cover.

Acknowledgements

The following people have been helpful in providing me with data, comments, references and ideas during this research:

- Jan van Dalen (RACM, Amersfoort)
- Boudewijn Goudswaard (Past2Present/Archeologic, Woerden)
- Dr. René Isarin (Past2Present/Archeologic, Woerden)
- Dr. Hans Kamermans (Faculty of Archaeology, Leiden University)
- Richard Kroes (Past2Present/Archeologic, Woerden)
- Dr. Martijn van Leusen (Institute of Archaeology, University of Groningen)
- Dr. Heleen van Londen (AAC Projectenbureau, University of Amsterdam)
- Prof. Clive Orton (Institute of Archaeology, University College London)
- Dr. Albertus Voorrips
- Milco Wansleeben (Faculty of Archaeology, Leiden University)

I would like to thank them all for their interest and their help. This research was made possible through a grant from NWO within the framework of the project ‘Strategic research into, and development of best practice for, predictive modelling on behalf of Dutch cultural resource management’. RAAP Archeologisch Adviesbureau provided additional funding and resources.

References

Altschul, J.H. 1988. Models and the Modelling Process. In W.J. Judge and L. Sebastian (eds), *Quantifying the Present and Predicting the Past: Theory, Method, and Application of Archaeological Predictive Modelling,* 61-96. Denver: U.S. Department of the Interior, Bureau of Land Management Service Center

Altschul, J.H. and C.R. Nagle 1988. Collecting New Data for the Purpose of Model Development, In W.J. Judge and L. Sebastian (eds), *Quantifying the Present and Predicting the Past: Theory, Method, and Application of Archaeological Predictive Modelling,* 257-300. Denver: U.S. Department of the Interior, Bureau of Land Management Service Center

Attema, P., G.-J. Burgers, E. van Joolen, M. van Leusen and B. Mater (eds) 2002. *New Developments in Italian Landscape Archaeology.* BAR International Series 1091. Oxford: Archaeopress

Atwell, M.R. and M. Fletcher 1985. A new technique for investigating spatial relationships: significance testing. In A. Voorrips and S.H. Loving (eds), *To pattern the past. Proceedings of the Symposium on Mathematical Methods in Archaeology, Amsterdam 1984 (PACT II),* 181-190. Strasbourg: Council of Europe

Atwell, M.R. and M. Fletcher 1987. An analytical technique for investigating spatial relationships. *Journal of Archaeological Science*,14, 1-11

Banko, G. 1998. A *review of Assessing the Accuracy of Classifications of Remotely Sensed Data and of Methods Including Remote Sensing Data in Forest Inventory. Interim Report IR-98-081.* Laxenburg: International Institute for Applied System Analysis
http://www.iiasa.ac.at/Publications/Documents/IR-98-081.pdf, accessed 25-01-2005

Banning, E.B. 2002. *Archaeological Survey.* Manuals in Archaeological Method, Theory and Technique. New York: Kluwer Academic / Plenum Publishers
Baxter, M.J. 2003. *Statistics in Archaeology.* London: Hodder Arnold

Bonham-Carter, G.F. 1994. *Geographic Information Systems for Geoscientists: Modelling with GIS.* Computer Methods in the Geosciences Volume 13. Pergamon

Brandt, R.W., B.J. Groenewoudt and K.L. Kvamme 1992. An experiment in archaeological site location: modelling in the Netherlands using GIS techniques. *World Archaeology* 24, 268-282

Buurman, J. 1996. *The eastern part of West-Friesland in Later Prehistory. Agricultural and environmental aspects.* PhD thesis, Rijksuniversiteit Leiden, Leiden

Carmichael, D.L. 1990. GIS predictive modelling of prehistoric site distribution in central Montana. In K.M.S. Allen, S.W. Green and E.B.W. Zubrow (eds), *Interpreting Space: GIS and Archaeology,* 216-225. New York: Taylor and Francis

Cochran, W.G. 1963. *Sampling techniques, second edition.* New York: John Wiley & Sons, Inc.

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20(1), 37-40

Dalla Bona, L. 1994. *Ontario Ministry of Natural Resources Archaeological Predictive Modelling Project*. Thunder Bay: Center for Archaeological Resource Prediction, Lakehead University

Dalla Bona, L. 2000. Protecting Cultural Resources through Forest Management Planning in Ontario Using Archaeological Predictive Modeling. In K.L. Wescott and R.J. Brandon (eds), *Practical Applications of GIS for Archaeologists. A Predictive Modeling Toolkit,* 73-99. London: Taylor and Francis

Deeben, J., D. Hallewas, J. Kolen, and R. Wiemer 1997. Beyond the crystal ball: predictive modelling as a tool in archaeological heritage management and occupation history. In W. Willems, H. Kars, and D. Hallewas (eds), *Archaeological Heritage Management in the Netherlands. Fifty Years State Service for Archaeological Investigations*, 76-118. Amersfoort: Rijksdienst voor het Oudheidkundig Bodemonderzoek

Deeben, J.H.C., D.P. Hallewas and Th.J. Maarlevelt 2002. Predictive modelling in archaeological heritage management of the Netherlands: the indicative map of archaeological values (2nd generation). *Berichten ROB* 45, 9-56. Amersfoort: Rijksdienst voor het Oudheidkundig Bodemonderzoek

Delicado, P. 1999. Statistics in Archaeology: New Directions. In J.A. Barceló, I. Briz and A. Vila (eds), *New Techniques for Old Time. CAA98 – Computer Applications and Quantitative Methods in Archaeology. Proceedings of the 26th Conference, Barcelona, March 1998*. BAR International Series 757, 29-38. Oxford: Archaeopress
www.econ.upf.edu/docs/papers/downloads/310.pdf

Ducke, B. and U. Münch 2005. Predictive Modelling and the Archaeological Heritage of Brandenburg (Germany) In M. van Leusen and H. Kamermans (eds), *Predictive Modelling for Archaeological Heritage Management: A research agenda*. Nederlandse Archeologische Rapporten 29, 93-108. Amersfoort: Rijksdienst voor het Oudheidkundig Bodemonderzoek

Duncan, R.B. and K.A. Beckman 2000. The Application of GIS Predictive Site Location Models within Pennsylvania and West Virginia. In K.L. Wescott and R.J. Brandon (eds), *Practical Applications of GIS for Archaeologists. A Predictive Modeling Toolkit,* 33-58. London: Taylor and Francis

Ebert, J.I. 2000. The State of the Art in "Inductive" Predictive Modeling: Seven Big Mistakes (and Lots of Smaller Ones). In K.L. Wescott and R.J. Brandon (eds), *Practical Applications of GIS for Archaeologists. A Predictive Modeling Toolkit,* 129-134. London: Taylor and Francis

Efron, B. 1979. Bootstrap methods: another look at the jackknife. *The Annals of Statistics* 7, 1-26

Efron, B. and R.J. Tibshirani 1993. *An Introduction to the Bootstrap*. Monographs on Statistics and Applied Probability 57. New York: Chapman & Hall

Ejstrud, B. 2003. Indicative Models in Landscape Management: Testing the Methods. In J. Kunow and J. Müller (eds), *Landschaftsarchäologie und geographische Informationssysteme: Prognosekarten, Besiedlungsdynamik und prähistorische Raumordnungen. 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, 119-134. Wünsdorf: Brandenburgisches Landesamt für Denkmalpflege und Archäologisches Landesmuseum

Ejstrud, B. 2005. Taphonomic Models: Using Dempster-Shafer theory to assess the quality of archaeological data and indicative models. In M. van Leusen and H. Kamermans (eds), *Predictive Modelling for Archaeological Heritage Management: A research agenda*. Nederlandse Archeologische Rapporten 29, 183-194. Amersfoort: Rijksdienst voor het Oudheidkundig Bodemonderzoek

Ente, P.J. 1963. *Een bodemkartering van het tuinbouwcentrum 'De Streek'.* De bodemkartering van Nederland, deel 21. Wageningen: Stiboka

Gibbon, G.E. 2002. *A Predictive Model of Precontact Archaeological Site Location for the State of Minnesota. Appendix A: Archaeological Predictive Modelling: An Overview.* Saint Paul: Minnesota Department of Transportation
http://www.mnmodel.dot.state.mn.us/chapters/app_a.htm, accessed on 25-01-2005

Gibbon, G.E., C.M. Johnson and S. Morris 2002. *A Predictive Model of Precontact Archaeological Site Location for the State of Minnesota. Chapter 5: The Archaeological Database.* Saint Paul: Minnesota Department of Transportation
http://www.mnmodel.dot.state.mn.us/chapters/chapter5.htm, accessed on 25-01-2005

Gibson, T.H. 2005. Off the Shelf: Modelling and management of historical resources. In M. van Leusen and H. Kamermans (eds), *Predictive Modelling for Archaeological Heritage Management: A research agenda*. Nederlandse Archeologische Rapporten 29, 205-223. Amersfoort: Rijksdienst voor het Oudheidkundig Bodemonderzoek

Griffin, D. and T.E Churchill 2000. 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), 137-153

Goodchild, M.F. 1986. *Spatial Autocorrelation*. CATMOG 47. Norwich: Geo Books
Hand, D.J. 1997. *Construction and Assessment of Classification Rules.* Chichester: John Wiley & Sons

Heunks, E. 2001. *Gemeente Ede; archeologische verwachtingskaart.* RAAP-rapport 654, Amsterdam: RAAP Archeologisch Adviesbureau

Heunks, E., D.H. de Jager and J.W.H.P. Verhagen 2003. *Toelichting Limes-kaart Gelderland; provincie Gelderland.* RAAP-rapport 860, RAAP Amsterdam: Archeologisch Adviesbureau

Hobbs, E. 2003. The Minnesota Archaeological Predictive Model. In J. Kunow and J. Müller (eds), *Landschaftsarchäologie und geographische Informationssysteme: Prognosekarten, Besiedlungsdynamik und prähistorische Raumordnungen. 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, 141-150. Wünsdorf: Brandenburgisches Landesamt für Denkmalpflege und Archäologisches Landesmuseum

Hobbs, E., C.M. Johson and G.E. Gibbon 2002. *A Predictive Model of Precontact Archaeological Site Location for the State of Minnesota. Chapter 7: Model Development and Evaluation.* Saint Paul: Minnesota Department of Transportation.
http://www.mnmodel.dot.state.mn.us/chapters/chapter7.htm, accessed on 25-01-2005

Hole, B. 1980. Sampling in archaeology: a critique. *Annual Review of Anthropology* 9, 217-234

Howson, C and P. Urbach 1993. *Scientific Reasoning: the Bayesian Approach. Second Edition*. Chicago: Open Court

Hudson, W. and Ramm, C. 1987. Correct formula of the Kappa coefficient of agreement. *Photogrammetric Engineering and Remote Sensing*, 53(4), 421-422

IJzereef, G.F. and J.F. van Regteren Altena 1991. Middle and Late Bronze Age settlements at Andijk and Bovenkarspel. In H. Fokkens and N. Roymans (eds), *Bronze Age and Early Iron Age settlements in the Low Countries*. Nederlandse Archeologische Rapporten 13, 61-81. Amersfoort: Rijksdienst voor het Oudheidkundig Bodemonderzoek

Jager, D.H. de, 1999. *PWN-transportleiding Hoorn-Andijk (deeltracé Wervershoof-Andijk), provincie Noord-Holland; archeologische begeleiding cultuurtechnisch onderzoek*. RAAP-rapport 440. Amsterdam: RAAP Archeologisch Adviesbureau

Kamermans, H. 2000. Land evaluation as predictive modelling: a deductive approach. In G. Lock (ed.), *Beyond the Map. Archaeology and Spatial Technologies.* NATO Science Series, Series A: Life Sciences, vol. 321, 124-146. Amsterdam: IOS Press / Ohmsha

Kamermans, H. 2003. Predictive Maps and Land Quality Mapping. In J. Kunow and J. Müller (eds), *Landschaftsarchäologie und geographische Informationssysteme: Prognosekarten, Besiedlungsdynamik und prähistorische Raumordnungen. 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, 151-160. Wünsdorf: Brandenburgisches Landesamt für Denkmalpflege und Archäologisches Landesmuseum

Kamermans, H., J. Deeben, D. Hallewas, P. Zoetbrood, M. van Leusen and Ph. Verhagen 2005. Project proposal. In M. van Leusen and H. Kamermans (eds), *Predictive Modelling for Archaeological Heritage Management: A research agenda*. Nederlandse Archeologische Rapporten 29, 13-23. Amersfoort: Rijksdienst voor het Oudheidkundig Bodemonderzoek

Kamermans, H. and E. Rensink 1999. GIS in Palaeolithic Archaeology. A Case Study from the Southern Netherlands. In L. Dingwall, S. Exon, V. Gaffney, S. Laflin and M. van Leusen (eds), *Archaeology in the Age of the Internet. Computer Applications and Quantitative Methods in Archaeology.* BAR International Series 750, 81 and CD-ROM. Oxford: Archaeopress

Kvamme, K.L. 1988a. Using existing data for model building. In W.J. Judge and L. Sebastian (eds), *Quantifying the Present and Predicting the Past: Theory, Method, and Application of Archaeological Predictive Modelling,* 301-324. Denver: U.S. Department of the Interior, Bureau of Land Management Service Center

Kvamme, K.L. 1988b. Development and Testing of Quantitative Models In W.J. Judge and L. Sebastian (eds), *Quantifying the Present and Predicting the Past: Theory, Method, and Application of Archaeological Predictive Modelling,* 325-428. Denver: U.S. Department of the Interior, Bureau of Land Management Service Center

Kvamme, K.L. 1990. 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, Vol. 3, 257-295. Bonn: Holos-Verlag

Kvamme, K. L. 1992. A Predictive Site Location Model on the High Plains: An Example with an Independent Test. *Plains Anthropologist,* 37, 19-40

Kvamme, K.L. 1993. Spatial Statistics and GIS: an integrated approach. In J. Andresen, T. Madsen and I. Scollar (eds), *Computing the Past. CAA92 – Computer Applications and Quantitative Methods in Archaeology,* 91-103. Aarhus: Aarhus University Press

Lange, S., S. Zijlstra, J. Flamman and H. van Londen 2000. *De archeologische begeleiding van het waterleidingtracé Andijk-West – Wervershoof.* Amsterdam: Amsterdams Archeologisch Centrum, Universiteit van Amsterdam

Lauwerier, R.C.G.M. and R.M. Lotte (eds), 2002. *Archeologiebalans 2002. Amersfoort:* Rijksdienst voor het Oudheidkundig Bodemonderzoek

Leusen, P.M. van 2002. *Pattern to Process: Methodological Investigations into the Formation and Interpretation of Spatial Patterns in Archeological Landschapes.* PhD thesis. Rijksuniversiteit Groningen, Groningen. http://www.ub.rug.nl/eldoc/dis/arts/p.m.van.leusen, accessed 03-06-2005

Leusen, P.M. van, J. Deeben, D. Hallewas, P. Zoetbrood, H. Kamermans and Ph. Verhagen 2005. A Baseline for Predictive Modelling in the Netherlands. In M. van Leusen and H. Kamermans (eds), *Predictive Modelling for Archaeological Heritage Management: A research agenda*. Nederlandse Archeologische Rapporten 29, 25-92. Amersfoort: Rijksdienst voor het Oudheidkundig Bodemonderzoek

Lunneborg, C.E. 2000. *Data Analysis by Resampling: Concepts and Applications.* Pacific Grove: Duxbury Press

Millard, A. 2005. What Can Bayesian Statistics Do For Predictive Modelling? In M. van Leusen and H. Kamermans (eds), *Predictive Modelling for Archaeological Heritage Management: A research agenda*. Nederlandse Archeologische Rapporten 29, 169-182. Amersfoort: Rijksdienst voor het Oudheidkundig Bodemonderzoek

Moon, H. 1993. *Archaeological Predictive Modelling: An Assessment*. RIC report 106. Resources Inventory Committee, Earth Sciences Task Force, Victoria: Ministry of Sustainable Resource Management, Province of British Columbia
srmwww.gov.bc.ca/risc/o_docs/culture/016/ric-016-07.htm, accessed 27-01-2005

Mooney, C. Z. and Duval, R. D. 1993. *Bootstrapping: A nonparametric approach to statistical inference*. Newbury Park: Sage Publications

Nance, J.D. 1983. Regional sampling in archaeological survey: the statistical perspective, In Schiffer, M.B. (ed.), *Advances in Archaeological Method and Theory* 6, 289-356. New York: Academic Press

Orton, C. 2000a. *Sampling in Archaeology.* Cambridge Manuals in Archaeology. Cambridge: Cambridge University Press

Orton, C. 2000b. 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, 1-7. Oxford: Archaeopress

Petrie, J. E. 1998. *The Accuracy of River Bed Sediment Samples*. MSc thesis. Blacksburg: Virginia Polytechnic Institute and State University
http://scholar.lib.vt.edu/theses/available/etd-011599-103221/unrestricted/Thesis.pdf, accessed 25-01-2005

Rose, M.R. and J.H. Altschul 1988. An Overview of Statistical Method and Theory for Quantitative Model Building. In W.J. Judge and L. Sebastian (eds), *Quantifying the Present and Predicting the Past: Theory, Method, and Application of Archaeological Predictive Modelling,* 173-256. Denver: U.S. Department of the Interior, Bureau of Land Management Service Center

Rosenfield, G. and Fitzpatrick-Lins, K. 1986. A coefficient of agreement as a measure of thematic classification accuracy. *Photogrammetric Engineering and Remote Sensing* 52(2), 223-227

Shennan, S. 1985. *Experiments in the Collection and Analysis of Archaeological Survey Data: The East Hampshire Survey.* Sheffield: Department of Archaeology and Prehistory, Sheffield University

Shennan, S. 1997. *Quantifying Archaeology. 2nd Edition*. Edinburgh: Edinburgh University Press

Simon, J.L. 1969. *Basic Research Methods in Social Sciences: the Art of Empirical Investigation*. New York: Random House

Simon, J.L. 1997. *Resampling: The New Statistics. 2nd Edition*
http://www.resample.com/content/text/index.shtml, accessed on 25-01-2005

Simon, J.L. 1998. *The philosophy and Practice of Resampling Statistics.* Unfinished manuscript.
www.resample.com/content/teaching/philosophy/index.shtml, accessed on 25-01-2005.
Smets, P. 1994. What is Dempster-Shafer's model?, In R.R. Yager, J. Kacprzyk and M. Fedrizzi (eds), *Advances in Dempster-Shafer Theory of Evidence,* 5-34. New York: Wiley.
http://iridia.ulb.ac.be/~psmets/WhatIsDS.pdf, accessed 20-10-2005.

Tol, A., Ph. Verhagen, A. Borsboom and M. Verbruggen 2004. *Prospectief boren. Een studie naar de betrouwbaarheid en toepasbaarheid van booronderzoek in de prospectiearcheologie.* RAAP-rapport 1000. Amsterdam: RAAP Archeologisch Adviesbureau

Verhagen, Ph. 2005. Prospecting Strategies and Archaeological Predictive Modelling. In M. van Leusen and H. Kamermans (eds), *Predictive Modelling for Archaeological Heritage Management: A research agenda*. Nederlandse Archeologische Rapporten 29, 109-121. Amersfoort: Rijksdienst voor het Oudheidkundig Bodemonderzoek

Verhagen, Ph. 2006. Quantifying the Qualified: the Use of Multi-Criteria Methods and Bayesian Statistics for the Development of Archaeological Predictive Models. In M. Mehrer and K. Wescott (eds), *GIS and Archaeological Predictive Modeling*, 191-216. Boca Raton: CRC Press

Verhagen, Ph. 2007. Predictive models put to the test In Ph. Verhagen, *Case Studies in Archaeological Predictive Modelling.* ASLU 14, 115-168. Leiden University Press

Verhagen, Ph. and J.-F. Berger 2001. The hidden reserve: predictive modelling of buried archaeological sites in the Tricastin-Valdaine region (Middle Rhône Valley, France). In Z. Stančič and 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, 219-232. Oxford: Archaeopress

Verhoeven, A.A.A. 1991. Visibility factors affecting artifact recovery in the Agro Pontino survey. In A. Voorrips, S.H. Loving, and H. Kamermans (eds), *The Agro Pontino Survey Project.* Studies in Prae- and Protohistorie 6, 87-98. Amsterdam: Instituut voor Pre- en Protohistorische Archeologie, Universiteit van Amsterdam

Wansleeben, M. and L.B.M. Verhart 1992. The Meuse Valley Project: GIS and site location statistics. *Analecta Praehistorica Leidensia* 25, 99-108

Warren, R.E. 1990a. Predictive modelling in archaeology: a primer. In K.M.S. Allen, S.W. Green and E.B.W. Zubrow (eds), *Interpreting Space: GIS and Archaeology*, 90-111. New York: Taylor and Francis

Warren, R.E. 1990b. Predictive Modelling of archaeological site location: a case study in the Midwest. In K.M.S. Allen, S.W. Green and E.B.W. Zubrow (eds), *Interpreting Space: GIS and Archaeology*, 201-215. New York: Taylor and Francis

Warren, R.E. and D.L. Asch 2000. A Predictive Model of Archaeological Site Location in the Eastern Prairie Peninsula. In K.L. Wescott and R.J. Brandon (eds), *Practical Applications of GIS For Archaeologists. A Predictive Modeling Kit,* 5-32. London: Taylor and Francis

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

Whitley, T.G. 2005a. A Brief Outline of Causality-Based Cognitive Archaeological Probabilistic Modeling. In M. van Leusen and H. Kamermans (eds), *Predictive Modelling for Archaeological Heritage Management: A research agenda*. Nederlandse Archeologische Rapporten 29, 123-137. Amersfoort: Rijksdienst voor het Oudheidkundig Bodemonderzoek

Whitley, T.G. 2005b. Re-thinking Accuracy and Precision in Predictive Modeling. Proceedings of 'Beyond the artifact - Digital interpretation of the past'. Paper presented at CAA 2004, 13-17 April, 2004, Prato

Zijverden, W.K. van and W.N.H. Laan 2005. Landscape reconstructions and predictive modeling in archaeological research, using a LIDAR based DEM and digital boring databases. *Workshop Archäologie und Computer 9.* Vienna, CD-ROM

9. Dealing with uncertainty in archaeological prediction[70]

Martijn van Leusen[71], Andrew R. Millard[72] and Benjamin Ducke[73]

9.1 Introduction[74]

One of the main characteristics of our reconstructions of past societies, and of archaeology in general, is the important role played by ignorance and uncertainty: there is little in archaeology that we know for certain, and much that we know little or nothing about. Such uncertainty (*e.g.* about the quality of the extant body of knowledge, or about the location of unknown resources) is also inherent to archaeological heritage management. In a previous study (van Leusen & Kamermans 2005) we established that a probabilistic or 'weight of evidence' approach should be better suited to creating predictive models than the more traditional approaches of the recent past, which have tended to ignore these uncertainties. The BBO project team therefore decided that 'reasoning with uncertainty' should be a major research theme for the second phase of the project.

Following the BBO call for 'action research' it was decided to explore the theme by creating a worked example based on a widely published Dutch predictive model, in such a way that the benefits of these new approaches could be made clear to non-technical readers as well as to those working in heritage management positions. Accordingly, this chapter has two aims:

1. to demonstrate that these approaches can result in useful models for CRM decision support (hopefully more useful than the existing traditional models), and
2. that concepts of expertise and uncertainty can be given a place in formal models without compromising on robustness and replicability, and hence can provide the link to concepts of (financial) *risk* employed by the people who use these models in the real world.

9.1.1 Approaches

Our 2005 study established that there are several approaches to *reasoning with uncertainty* that could conceivably be explored in the course of a worked example, such as Bayesian inference, fuzzy logic, Partial Probability theory and Dempster-Shafer theory. Two of these will be explored here. In section 9.2 Andrew Millard demonstrates how prior and conditional probabilities obtained from a regional archaeological expert (Groenewoudt) can be used to calculate, through Bayesian inference techniques, the posterior probabilities of site presence within the study area:

... it is impossible to separate opinions (prior beliefs), data and decisions/actions. In the 'classical' approach, our opinions influence our procedures in all sorts of subtle and little-understood ways, for example in choosing the significance level of a hypothesis test. It's better to be as explicit as we can about our prior beliefs, and let the theory take care of how they interact with data to produce posterior beliefs, rather than to let them lurk at the backs of our minds and cloud a supposedly 'objective' belief. This way the Bayesian approach can be more than just a nice piece of mathematics. (Orton, 2003)

---

70 This chapter is based on an NWO-BBO funded workshop, held at Amsterdam on 17-21 January 2005. We would like to acknowledge the contributions of those present (other than the authors themselves): Bert Groenewoudt, Huub Scholte Lubberink and Roy van Beek participated as archaeological experts and guinea pigs; Hans Kamermans and Philip Verhagen helped 'translate' between the archaeological and methodological experts. We also want to express our gratitude to our host, director Marten Verbruggen of RAAP Archaeological Consultancy.

71 Rijksuniversteit Groningen (p.m.van.leusen@rug.nl).

72 University of Durham (a.r.millard@durham.ac.uk).

73 Oxford Archaeological Unit Limited (benjamin.ducke@oxfordarch.co.uk)

74 The figures from this chapter can be found in colour at http://www.aup.nl/do.php?a=show_visitor_book&isbn=9789087280673

Bayesian statistics differs from classical statistics in allowing the explicit incorporation of subjective prior beliefs into our statistical analysis. The Baseline Report (Van Leusen *et al*. 2005, 66) sets out the main reason for using Bayesian approaches, that is incorporation of expert prior knowledge in a formal and transparent way into predictive models, thus making them more rigorous and of higher quality.

In section 9.3 Benjamin Ducke shows how initial uncertainties and beliefs can be used as inputs for a predictive model using Dempster-Shafer theory (DST), and how outputs such as the 'plausibility map' can be used to quantify the developers' risk as well as to target further research. DST provides a less rigid framework for archaeological inference than does the Bayesian approach. As defined by Dempster (1967) and Shafer (1976), it is built around the concept of *belief*, which is a somewhat more relaxed, generalized version of mathematical probability. The theory's hallmark is its capability to explicitly represent uncertainty, allowing critical decision support for large scale planning in cultural resource management (CRM) mitigation processes.

9.1.2 The worked example: rijssen-wierden 'de borkeld'

For the choice of data set, we figured it was important to be able to compare any new models made by us, with existing models made earlier for the same area. These conditions are met by the dataset used for the first predictive model in the Netherlands, the Rijssen-Wierden area in the eastern Netherlands (Ankum & Groenewoudt 1990; figure 9.1) which was later used for a revised model by Brandt and others (1992) as well as for three generations of IKAW. It is worth quoting the relevant paragraph from the Baseline Report:

RAAP became involved in archaeological zonation in the late 1980s (Brandt 1987). Initially the 'classical' American inductive method, using multivariate statistics, was advocated (Ankum & Groenewoudt 1990; Brandt 1990). A pilot study carried out in the Pleistocene landscape of the Rijssen-Wierden area resulted in a predictive model based on six environmental variables: soil substrate, landscape unit, zonal area, distance to water, distance to macro-gradient, and distance to water or macro-gradient. This methodology was subsequently applied again only once (in the regional study 'Ede'; Soonius & Ankum 1991a,b), when six rather different landscape variables were found to be significant in influencing archaeological prediction: soil type, geomorphology, ground water level, distance to macro-gradient, distance to surface water, and distance to dry valleys. Two additional factors quantifying geographical complexity were investigated as well, but were not found to be significant predictors (Soonius & Ankum 1991b: 72-74): heterogeneity or the number of distinct legend units of the soil and geomorphological maps occurring within a radius of 500m and 1000m, and variation or the number of zones occurring within that radius. Also new was the specification of predictions for five periods (late Palaeolithic and Mesolithic, Neolithic, Bronze Age, Iron Age and Roman period, and Middle Ages), where the Rijssen-Wierden model had lumped all periods together except for the Late Middle Ages - which was omitted altogether because it barely correlated with characteristics of the physical environment. A 'map of archaeological potential' was produced by factoring in both zones of disturbed soil and zones where circumstances were especially likely to conserve archaeological remains. The end result was then generalised for publication (van Leusen *et al*. 2005, 36-37).

We found one of the original researchers (Bert Groenewoudt, then with the RACM) prepared to help out with the data and to act as regional expert. The original GIS files for this area, luckily still kept at RAAP, were sent to Millard and Ducke along with a documentation of the map layers[75]. Roy van Beek, at the time doing PhD research on the eastern Netherlands at the RACM, was brought in as the second regional expert. For the purposes of comparison, we have also approached Huub Scholte Lubberink, then at RAAP Oost, to act as our third regional expert.

[75] In GRASS 4.1 format.

Figure 9.1 Location of the Rijssen-Wierden area (1) in the Netherlands.

The landscape in the 120 sq km area covered by sheet 28D of the Topographical Map of the Netherlands is characterized by glacial and periglacial forms which, because of their poor drainage, were later dotted with marshes and partly covered by peats. The area is crossed by several valleys with local clayey deposits. Wind-blown sandy areas were fixed by planted woods in the 19th century; near the present town of Rijssen we also find man-made soils covering the original landscape.

The relation landscape – prehistoric settlement was investigated (for the Pleistocene sandy areas only) by comparing the characteristics of 76 archaeological settlement or burial sites to those of 80 random control points, using a 100m raster resolution. The archaeological sites were subdivided by period into five groups (15 Palaeolithic and Mesolithic; 18 Neolithic to Middle Bronze Age; 16 Late Bronze Age to Iron Age; 8 Roman to Early Medieval Period, and 19 Late Medieval Period). Significant correlations were found to exist between sites and the following eight environmental variables: soil type, (geomorphological) landscape unit, soil substrate[76], groundwater table, soil/morphological unit size[77], distance to macrogradient[78], distance to surface water[79], and (probably) degree of heterogeneity[80]. All variables were derived from the following three map sources: sheet 28D of the topographic map 1:25,000, sheet 28 West of the soil map 1:50,000, and sheet 28/29 of the geomorphological map 1:50,000. No significant correlations could be found with three other variables (distance to boundary of low/wet areas, aspect, and elevation), probably because of the relatively subdued relief[81].

Ankum and Groenewoudt's predictive model used only those five factors which proved to be most useful in the study area, namely: soil substrate, landscape unit, unit size, distance to water, and distance to gradient[82]. The 'factor maps' for these variables were reclassified into two to four rated zones depending on the

[76] The soil *substrate* is likely to be the same as in the remote past, whereas the soil *type* may well have changed in the meantime.

[77] Some units, although suitable in other respects, might prove to be unsuitable for agricultural use because they are too small.

[78] This is based on the idea that the physically most diverse parts of the landscape are also ecologically the richest, hence the most attractive in some periods. The boundaries between the high glacial and sandy area and the low-lying area are one example of this.

[79] For both drinking water and hunting/fishing.

[80] As measured by the number of soil/geomorphological polygons within 0.5, 1, or 2 km from the site, and the number of different soil/geomorphological legend units within 0.5, 1, or 2 km from the site.

[81] Correlations were, however, significant for subgroups of sites: see Ducke's analysis in section 9.3.5.

[82] Note that, although the site-environment correlations were investigated for each period separately, the choice for these five factors is determined by the authors' attempt to pick the 'overall best' predictors.

positive or negative influence they seemed to exert over site location choice; then all factor maps were added together to obtain a single raster map in which higher values signify greater site potential. The resulting value range, with a minimum of 0 and a maximum of 13 points, was reclassified into the range 1 - 4 to arrive at the final archaeological potential map (figure 9.2). This model attempted to optimize *gain*[83], and in fact placed 73% of the input sites into the two top classes which together make up 24% of the area, for a gain of 67%. The authors note that, if Late Medieval settlements (which do not correlate well with the physical landscape) had been left out of the model, the gain would probably have been higher. Conducting field checks in areas with a positive expectation value only, the authors subsequently located another 27 new archaeological sites.

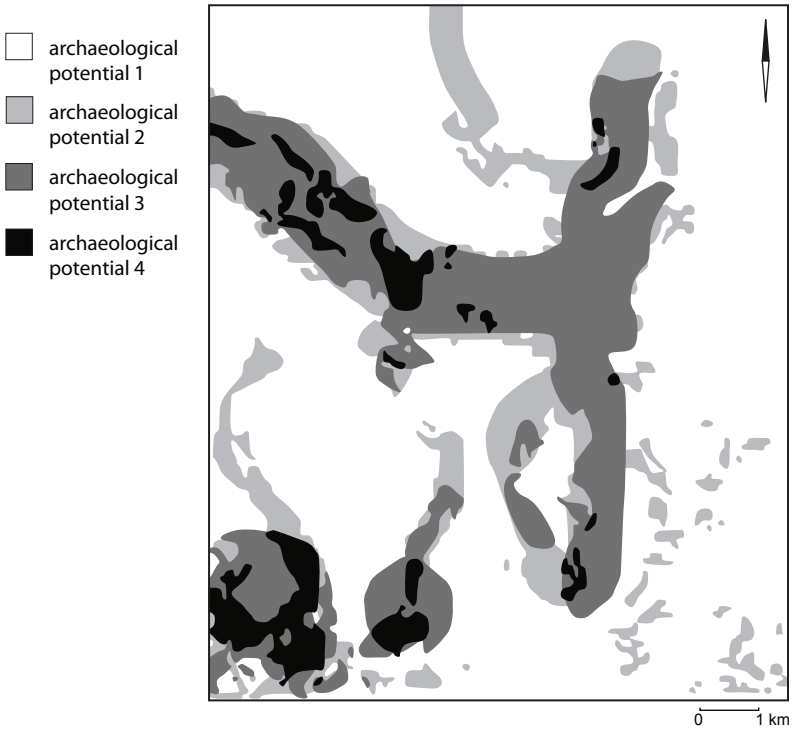

Figure 9.2 Ankum & Groenewoudt's (1990, figure 13) predictive model for Rijssen-Wierden. The four shades of gray represent zones of increasing archaeological potential.

Re-building the original data set and derived models, as depicted in Appendices 24-29 and figures 13 and 14 of Ankum and Groenewoudt's 1990 study, with full fidelity proved to be impossible because the data themselves and/or the procedures used to produce the maps were not always adequately described. However, we succeeded eventually in obtaining raster layers that closely resemble the 1990 'factor maps' (depicted in figures 9.12-9.16).

In addition to the 79 sites used in the original 1990 study and the 50 new sites used for the 1991 test, Groenewoudt provided a further 17 site locations as well as 23 'non-site' locations discovered since 1991[84] and Scholte Lubberink provided an additional list of 100 probable Late Medieval (LMED) sites[85].

[83] Calculated here as 1 - %area / %sites.

[84] Observed 'non-site' locations, *e.g.* fields that were surveyed under good visibility conditions and where no archaeology was observed, can under certain conditions be used to estimate the probability of a site occurring in a particular landscape setting. See the discussion in section 9.4.

[85] Based on the locations of post-Medieval toponyms indicating the presence of farmsteads; LMED sites are thought to be located within 150 m of these locations.

9.1.3 Tapping the expertise

Besides data representing the physical landscape and the locations of the known archaeological sites, we needed to obtain information about both location factors and bias factors from the regional archaeological experts in a structured manner.

First, we asked each of them to independently weigh all categories distinguished in the reconstructed 1990 factor maps (Ankum and Groenewoudt 1990, appendices 24 – 29), so that we could demonstrate how such weights can be used to create a Bayesian model of prior probabilities. However, the experts believed that more realistic results could be achieved by changing the categories in some of the factor maps and by distinguishing several periods rather than lumping everything together. Hence the maps representing soil substrate and geomorphology were reclassified by the experts jointly, before being weighted again independently. Likewise, the maps representing polygon size and distance-to-water received different weightings for the four broad periods that the experts considered the most relevant (Palaeolithic and Mesolithic (PAL-MES), Neolithic to Early Bronze Age (NEO-EBA), Middle Bronze Age to Early Medieval Period (MBA-EMED), and Late Medieval Period (LMED))[86].

Expertise about the effects of *extraneous* factors (biases) on the discovery of archaeological remains was also modelled. The experts each assessed the relative probabilities of a site being discovered in each of six land-use types (Urban areas, Woodland, Moor, Pasture, Arable, and Water), always supposing a site were present. They also provided estimates of the percentage of sites already discovered within the study area, and estimates of the total numbers of sites present in each of the four broad periods. Groenewoudt further estimated the relative probabilities of discovery for each period.

One remarkable aspect of the Rijssen-Wierden dataset, demonstrating the importance of expert knowledge in predictive modeling, is the absence of sites in the built-up areas of Rijssen whereas many sites are present in the built-up area of Enter/Wierden[87]. The experts explained this as a consequence of a difference both in the level of amateur activity, and a difference in the chronology of local urban development: as development at Enter took place relatively late, better records were kept of archaeological discoveries. The upshot of this is that the experts *disbelieve* the absence of sites in Rijssen.

Finally, all three experts also supplied expertise on the effects of land use/land cover on research intensity, which sometimes proved to be counter-intuitive. On the one hand, arable land and cover sands had been intensively surveyed whereas the low-lying (generally grassy) areas were surveyed much less intensively; on the other, archaeological observations carried out during mechanized sod stripping allowed a high discovery rate in some heath areas.

9.2 A bayesian model demonstrating the use of expert opinion and the generation of uncertainty maps

Millard produced a multi-period Bayesian model for settlement in the study area, showing how conditional probabilities combine with observations to yield posterior probabilities, with an associated measure of uncertainty. In this section we explain, by example, the various steps in the model calculations, beginning with the quantification of expert opinions (‘obtaining prior proportions’), their conversion to equivalent numbers of sites and maps of expected relative site densities, the effect of adding actual site observations, and the use of maps depicting the amount of disagreement (uncertainty) among experts before and after observations were added.

---

[86] One technical option, which was not further explored by us, is the use of functions rather than categorical weights to describe a factor.

[87] This leads to a *conflict of evidence*, in which a particular factor appears to be both in favour of, and against, settlement.

Step 1: Obtaining prior proportions

Experts' relative odds of sites being present in each of the classes for each of the six factors were given for the 1990/1992 model, and for a new 2005 model. For example, expert's 2 odds with regard to the factor soil texture are given in table 9.1.

	a	b	c	d		a	b	c	d	sum
a	1	0,5	0,25	0,1		0,059	0,118	0,235	0,588	1
b	2	1	0,25	0,5		0,067	0,133	0,533	0,267	1
c	4	4	1	0,33		0,056	0,056	0,222	0,667	1
d	10	2	3	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 9.1 One of the three experts' assessment of relative odds for factor 'soil texture' categories a-d (left), converted into probabilities (right), with means and variances (bottom). CV= coefficient of variance.

On the left, each combination of texture classes is evaluated. For example, class (row) c is judged to be 4 times more likely to contain sites as is class (column) b. These relative odds supplied by the experts can be rewritten as absolute probabilities (known as 'prior proportions') summing to 1. Since the expert in question supplied information on all possible combinations of texture classes (in principle one data row would have been enough), it is possible to make four separate calculations of these prior proportions (the four rows in the right-hand part of the table), which need not (indeed, do not) agree. Continuing our example, texture class c according to this expert should 'attract' between 17 and 53 percent of the sites (column outlined), with a mean of 29% and a coefficient of variance of 56%. In other words, this expert is quite uncertain about some of his odds. The calculated means provide our current best estimate of this expert's true opinion on the importance of the four soil texture classes, so this is what we will use in further calculations.

Step 2: Calculating Dirichlet prior vectors

18 identical calculations were made for each of the six location factors and for each of the three experts separately (they were asked not to confer), and this information was combined to arrive at an assessment of the mean expert opinion and its variance, and the consequent 'data equivalent', which expresses our reliance on the experts' opinions in terms of the number of actual site observations that would be needed to dislodge them. For example, table 9.2 contains each of the three experts' prior proportions for the factor soil texture, along with the means and standard deviations. As can be seen, the three experts in this case are in rough agreement about the proportions of sites they would predict to lie in each of the four soil texture classes (if these were equally represented in the study area). For soil texture type c, for instance, expert opinions of 22, 29, and 30 percent average out at 27%, with a standard deviation of only 5%. There is somewhat less agreement on the 'attractiveness' of soil texture type d, with a standard deviation of 11%.

From this, so-called Dirichlet prior vectors and data equivalents can be calculated using various approaches (methods A to C in table 9.3). For example, in method A it is assumed that each expert is worth one observation – *i.e.*, the combined 'data equivalent' of our experts is 3. However, rather than making assumptions, we want the data equivalent to follow from the mean and variance of the experts' priors, and this is done in methods B1 and B2. The value α_0 is the apparent data-equivalent from the mean and variance of the experts' opinions for each soil texture class. Method B1 uses the mean of these α_0 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_0 values. Thus, for the factor soil texture, the experts' opinions are calculated to be worth between 17 (conservative method B2) and 39 (mean method B1) observations.

	a	b	c	d	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	0,300	0,500	1
mean	0,079	0,081	0,269	0,571	
SD	0,064	0,052	0,046	0,115	

Table 9.2 Experts' mean prior proportions for the factor 'soil texture', with means and standard deviations (SD).

When we repeat this calculation for all of the six location factors, then take the overall average data equivalent, the conservative calculation results in a value of 28 and the mean calculation in a value of 40. An overall data equivalent of 30 therefore seems to be a reasonable value to work with. In other words, in answer to our question of *how much weight should we assign to expert opinion* as opposed to actual observations, the answer seems to be that our expert's opinions are worth the equivalent of 30 observations. For the factor soil texture this translates (row C of table 9.3) to 2.4 observations in soil texture class a, 2.4 in class b, 8.1 in class c, and 17.1 in class d.

class method	a	b	c	d	data equivalent
A	0,24	0,24	0,81	1,71	3
α_0	16,82	26,11	94,04	17,44	
B1	3,03	3,12	10,39	22,05	38,60
B2	1,32	1,36	4,53	9,61	16,82
C	2,36	2,42	8,08	17,14	30

Table 9.3 Calculation of the experts' data equivalent for the factor soil texture, using Dirichlet prior vectors. Method A uses a prior data equivalent of 3; Method B1 uses the mean α_0 from the variance of expert opinions; Method B2 uses the minimum α_0 from the variance of expert opinions; and Method C uses a prior data equivalent of 30.

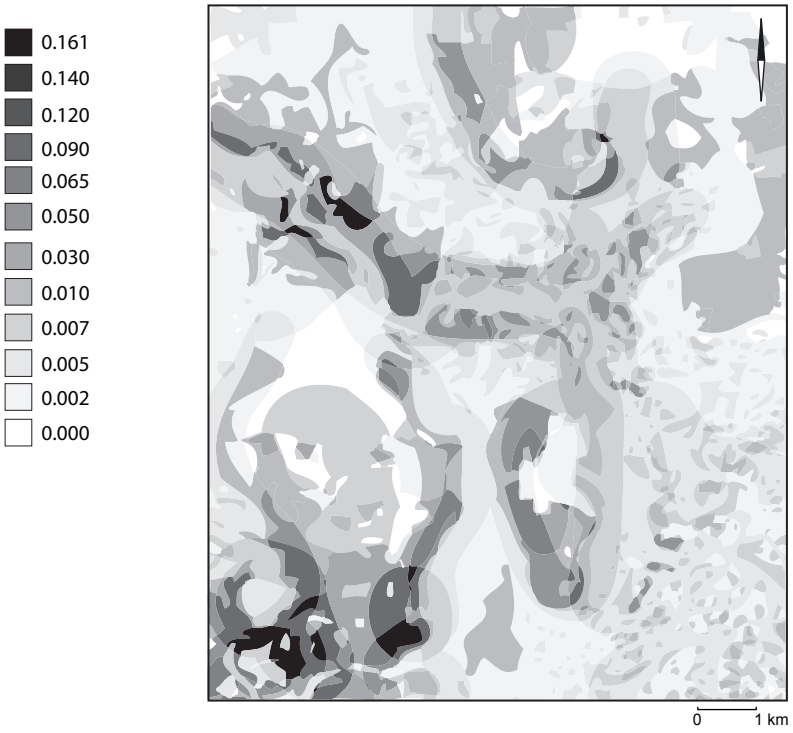

Figure 9.3 Relative site density according to expert judgement (prior proportions; a cell with a value of 0.12 is twice as likely to contain a site as a cell with a value of 0.06).

Where the experts agree (that is, where the variance of their opinions is low) a high data equivalent results; where they disagree a low one results. This is correct, since we do not value very highly the opinion of experts who cannot agree among themselves, whereas we tend to trust experts who find themselves in agreement[88]. Since the case study uses 80 actual sites, expert opinion may be said to represent about a quarter (30/110) of the final prediction in this particular model.

Step 3: Mapping expert opinion and confronting it with data

In order to generate a predictive model that uses this calculated 'weight of expert opninion' in combination with the circa 80 site observations available in 1992, a map containing all possible combinations of the 6 location factors was generated. There are 96 classes in this map: 4 texture classes * 3 geomorphological units * 2 unitsurface classes * 2 gradient distance classes * 2 water distance classes[89]. Using the Dirichlet prior vectors generated by method C, the relative probability of finding a site in each of these was calculated and mapped (figure 9.3). The map may be viewed as a summary of our experts' views on the relative density of sites in the landscape.

[88] Interestingly, when the 2005 model was built, the experts had already been conferring about the best way of representing location factors, and so their opinions had converged; resulting in dramatically lower measure of variance and, in consequence, a higher 'data equivalent': 700 – 1400 rather than 20 - 40. In effect, the three experts had reached agreement on the odds they would assign to each soil substrate and geomorphological class, hence their combined priors can only be dislodged by a very high number of contrary observations.

[89] Geomorphology classes 1 and 3, and 4 and 5, were combined for this part of the calculation, following Brandt *et al.*'s (1992) method as reconstructed by us.

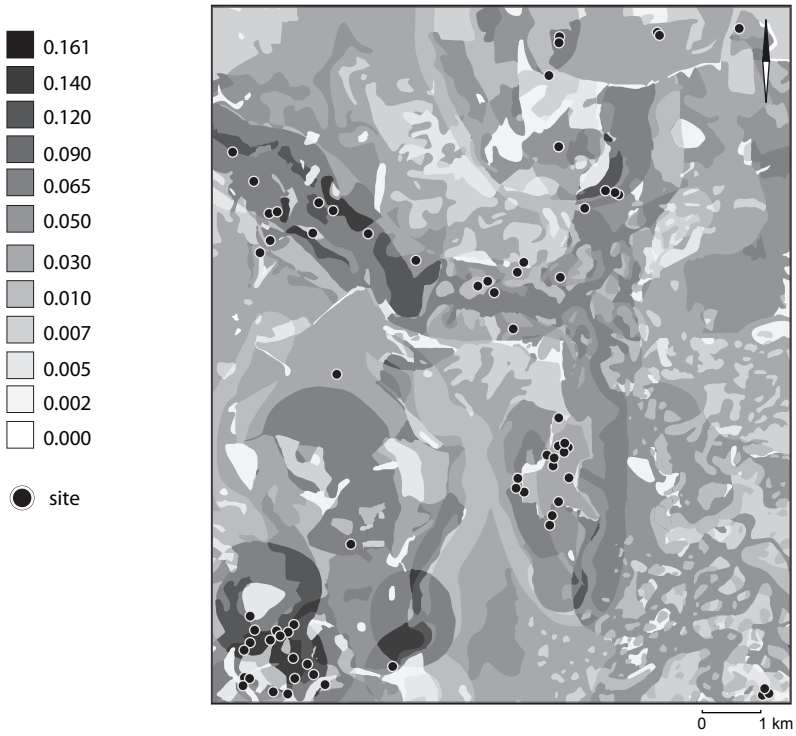

Figure 9.4 Relative site densities following inclusion of 80 observed sites (posterior proportions, using the same legend as figure 9.3) with sites overlaid.

This prediction can now be confronted with the 80 site observations available in 1992, and a number of discrepancies will be noted. We see areas where sites are present despite their predicted absence, as well as areas where sites are absent despite their predicted high density. This is partly as it should be: actual site discoveries are steered largely 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 exist, 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' data equivalent, and re-run the model calculations, a map of *posterior proportions* results (figure 9.4). If we compare this map to the prior proportions in figure 9.3, it will be seen that the addition of the sites has caused the model to somewhat increase its predicted site probability for most of the lighter areas (turning them grey in figure 9.4), and to generalize somewhat its predictions of zones of relatively high site density (dark shades in figure 9.4). In other words, the *posterior* probabilities have changed with respect to the *prior* probabilities under the influence of the observations, as per Bayes' theorem. The changes themselves can also be mapped (figure 9.5) for further study.

The case study thus far 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 trusted experts, 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.

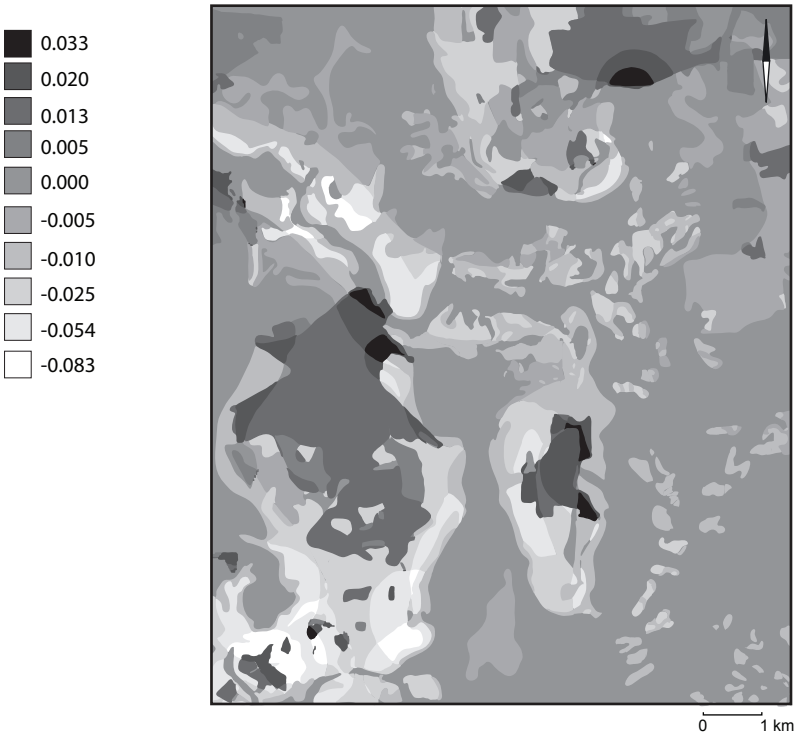

Figure 9.5 Difference between prior and posterior predicted site densities. Predicted site density increases (dark) or decreases (light) when observed sites are included into the model.

Happily, we do not need to be completely subjective in our rating of the quality of our experts: the variation among the expert's opinions itself provides us with a quantified measure of uncertainty, which itself can also be expressed as a map (figure 9.6). After again including the observed sites in the model, model uncertainties are shown to change (figure 9.7); the difference is depicted in figure 9.8. In this latter map, light grey indicate that we should have a high confidence in the predictive model, dark grey that we should be less confident that our model is correct. Such a map has obvious implications for the regional heritage management agenda: studies can be targeted at the high uncertainty zones in order to increase overall confidence in the predictive model.

Figures 9.6 and 9.7 show absolute SDs. The difference between them as shown in figure 9.8 is difficult to grasp because the relative uncertainty also depends on the mean relative site density. The difference becomes clearer if we calculate the SD as a fraction of the mean site density; once mapped (figure 9.9), this shows more clearly that uncertainty has decreased in all parts of the map relative to the priors.

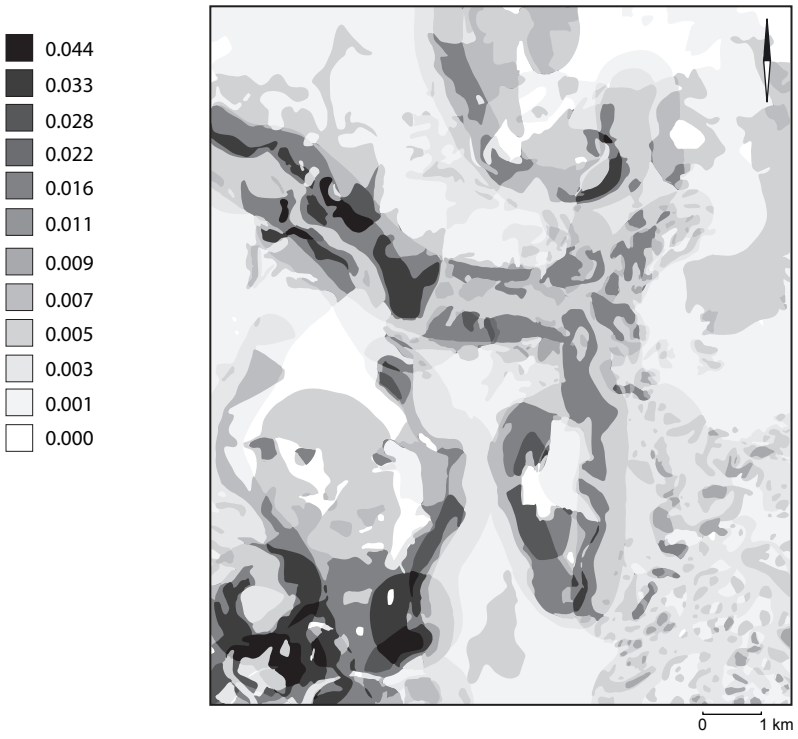

Figure 9.6 Uncertainty in the relative densities of sites as modelled by the experts (prior SD; darker shades indicate areas of greatest uncertainty. These correspond, unsurprisingly, to the areas of highest density in Figure 9.3.

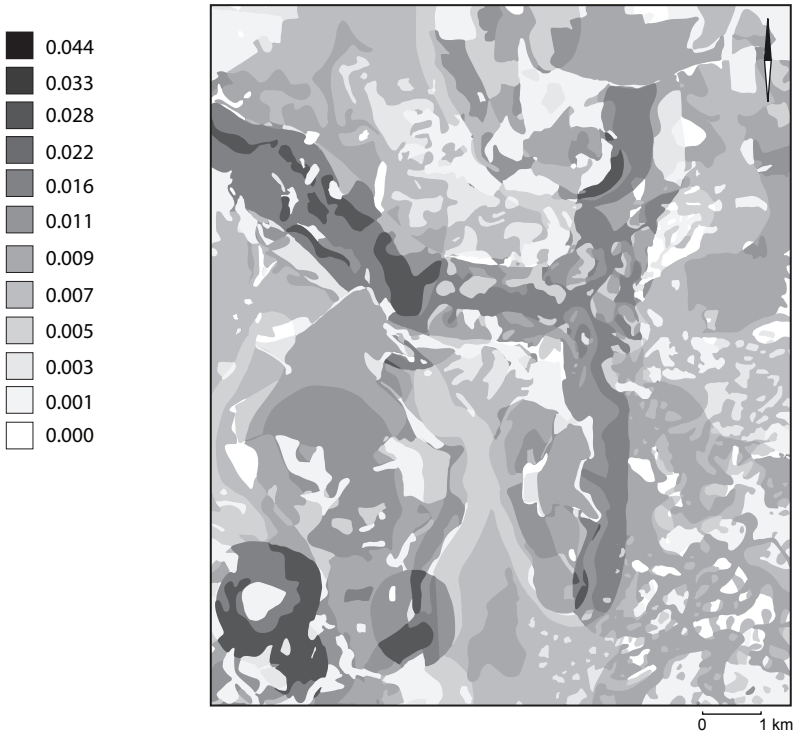

Figure 9.7 Uncertainty in the relative densities of sites after the data is included (posterior SD).

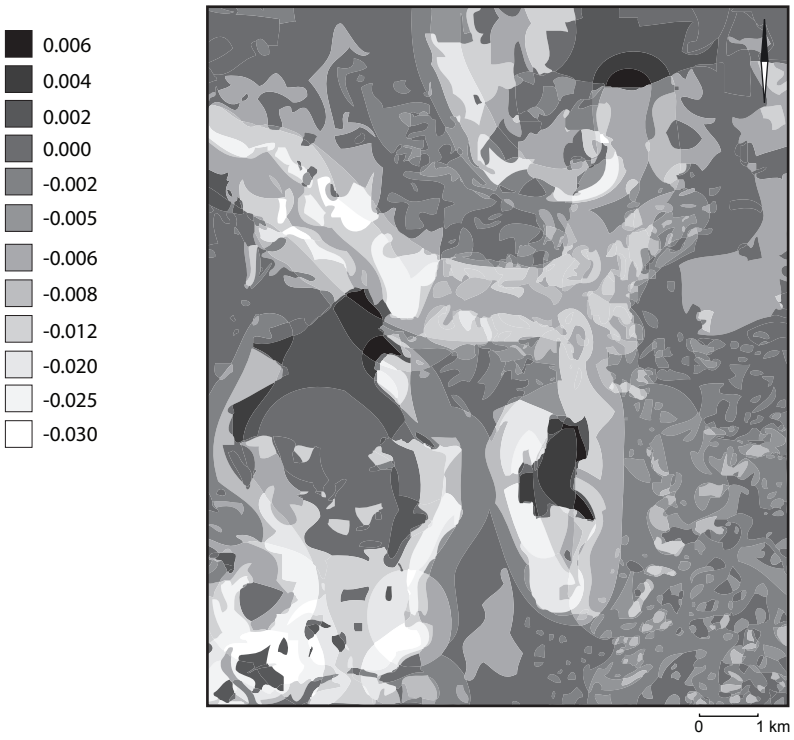

Figure 9.8 Difference between 9.6 and 9.7. Uncertainty has decreased in some areas (light) but increased in others (dark).

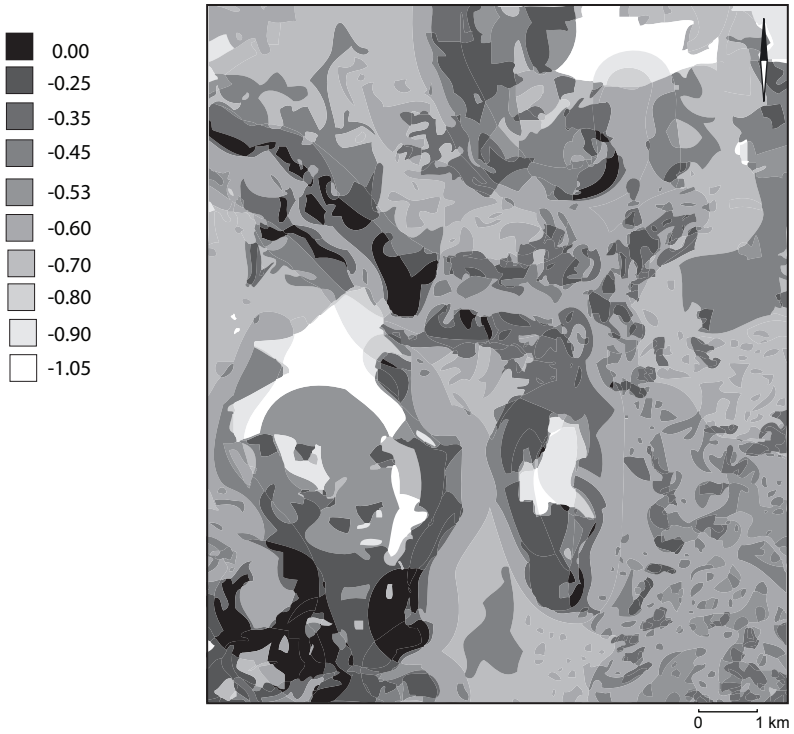

Figure 9.9 Change in uncertainty between prior and posterior, expressed as difference of relative SDs. Lighter shades signify greater reductions in uncertainty.

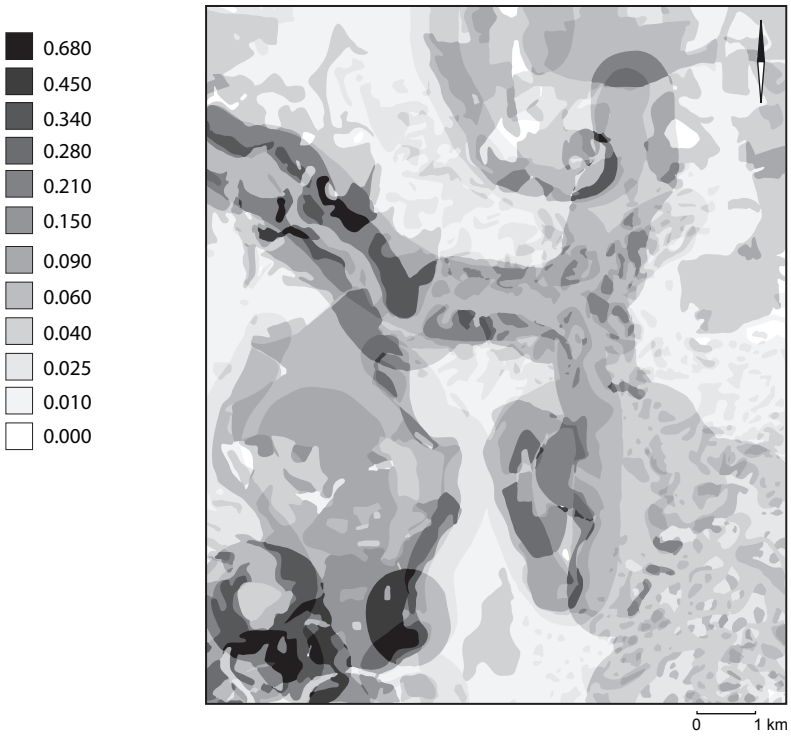

Figure 9.10 Posterior predicted total site density in numbers of sites per hectare. In the very highest density areas, a site is expected in every 1.5 hectares (68 sites per km2), but on average the density is only 0.06 sites per hectare (1 site per 16 ha).

Step 4: From relative to absolute site densities

A further advantage of our approach over the IKAW approach is that the resulting predictive model is quantified to a much greater extent: using experts' estimates of the total number of *undiscovered* sites in the study area, the model can provide estimated absolute site densities for each landscape class[90].

Since there are large differences in the experts' estimates of the total site count as well as in their distribution over the four periods, the standard deviations are substantial, and our confidence in the mean values therefore remains low. The experts stressed that they not have great confidence in their own capability to estimate these numbers reliably, and this is reflected in the differences between them as summarised in table 9.4.

Nonetheless we can now use the mean estimate of total number of sites, 755, to generate a map of predicted site density in numbers of sites per hectare (figure 9.10). Whilst the values in this map could be as much as 50% off[91], they could form the basis for cost calculations once the area affected by a planned building activity is known, and is therefore potentially of great use to government archaeologists and developers alike (cf. Gibson 2005). This kind of map could be field tested in areas of large-scale topsoil stripping, and subsequently corrected, but further study would be required to develop this type of application.

[90] The IKAW only provides relative average site densities for each of the three main prediction zones.

[91] Given the standard deviation of 336 sites on a mean of 755. Unpublished field data from the German Land Brandenburg indicate that the total site numbers estimated by experts may be far too low in general (pers. comm. Benjamin Ducke).

	estimated % of sites discovered	estimated no. of sites per period				estimated total no. of sites
		PAL-MES	NEO-EBA	MBA-EMED	LMED	
expert 1	10%	200	300	400	150	1050
expert 2	25%	200	30	60	100	390
expert 3	20%	300	100	175	250	825
mean	18,3%	233	143	212	167	755
SD	7,6%	58	140	173	76	336

Table 9.4 Summary of experts' opinions about the proportion of discovered versus undiscovered sites, and the distribution of all sites over the four periods distinguished.

9.3 A DEMPSTER-SHAFER MODEL SHOWING EXPLICIT UNCERTAINTY HANDLING AND UPDATING

Dempster-Shafer Theory (DST) is a theory of uncertainty. It is mathematically related to both set theory and probability theory, but provides a more flexible framework that has many interesting properties when it comes to handling uncertainty. Different applications and research interests have focused on different aspects of DST and keep producing new interpretations (see Smets 1994). Whilst this has led to different proposals for how to *calculate* DST models (Ejstrud 2005, 184), DST as used here follows Shafer's (1976) original proposal of a well defined, reasonably simple mathematical tool that fits a wide range of research applications. The aim in this section is to demonstrate how a DST approach can improve the practical uses and the quality of archaeological predictive models. The mathematical framework is further described in Appendix 1 for readers interested in the mathematical background.

Before we turn our attention to the case study itself, we must introduce some aspects of DST as used in an archaeological predictive modelling context. The first archaeological predictive model using DST was published by Ejstrud (2003; 2005)[92]. Ejstrud demonstrated the superiority of DST over various other predictive modelling approaches in terms of model performance, but it also has some conceptual benefits:

- simplicity: a DST predictive model can be set up in just a few, always identical steps; updating an existing model with new information is just as easy.
- flexibility: DST has the ability to combine information from different sources, regardless of distribution parameters.
- interpretability: the DST functions produce output that tends to be more meaningful than most statistical measures.
- realism: handling uncertainty explicitly produces decision support maps that do not assume complete information.

Translating our predictive modelling problem into the language of DST is a straightforward procedure:

- the *Frame of Discernment* (FoD) is the exhaustive set of hypotheses on the outcomes of the model (see below).
- a GIS map that encodes a variable with relevance to the FoD is a *source of evidence*.
- the entirety of GIS maps provided constitutes the body of evidence.
- each map object is transformed into an *evidence* by calculating a Basic Probability Number (BPN) *for it*. A BPN is the basic quantification of evidence in the DST. It expresses the strength of belief in the truth of a hypothesis in the light of a single source of evidence.

[92] Ejstrud points out, however, that the IDRISI software used for his research actually contains an archaeological scenario in the manual for its DST modelling tools.

In the case discussed here, the FoD is taken to consist of h_1={"site"}, which proposes that an archaeological site is present, h_2={"no site"}, which proposes that no archaeological site is present and $\{h_1,h_2\}$={"site","no site"}, which proposes that no decision can be made about site presence or absence. As demanded by DST, h_1 and h_2 are mutually exclusive and exhaustive, i.e. they cover all possible outcomes. The union of both $\{h_1,h_2\}$ represents uncertainty and we will refer to this as the "uncertainty hypothesis"[93].

9.3.1 Handling uncertainty

The greatest merit of DST-based predictive modelling is its capability to explicitly handle uncertainty. In predictive modelling, uncertainty often arises because there is direct evidence for the *presence* of a site, but only indirect evidence for the *absence* of one. Thus, the fact that no prehistoric sites were ever found on terrain type X might mean that (a) prehistoric settlers actually avoided this type of terrain, (b) the terrain type was in fact suitable but remained unused, or (c) one of many possible bias factors prevented the detection of the prehistoric sites that are actually present on this type of terrain. In cases like these we might not be able to decide between the "site" and "no site" hypotheses, and this inability to decide is the very nature of uncertainty.

DST gives us the possibility to transfer a certain amount of the basic probability mass (called *uncommitted* mass) to the uncertainty hypothesis {"site","no site"} and thus allows us to postpone decisions about "no site" until better information becomes available. Basic probability mass is transferred to the uncertainty hypothesis four cases: (1) we are not certain about the significance of a specific category (legend item) in one of the evidences (input maps), *e.g.* the soil type category 'built-up areas'; (2) we are not certain about the significance of an evidence, *e.g.* 'distance to surface waters'; (3) we are not certain that sites are *really* absent from specific map categories, even though we have not found them; and (4) we are not certain that a particular location has *correctly* been identified as a site. Here is a more technical decription of these four possibilities:

1. the probability *P* that the observed difference in proportion between sample (sites) and population (entire region) for evidence category *C* could also be produced by chance is greater > 0. In this case, *P* is subtracted from the mass for either "site" or "no site" and transferred to {"site","no site"} for category *C*. This leaves $m(h_1)$ or $m(h_2)$ as 1-*P*, respectively.
2. 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 mass for either "site" or "no site" and transferred to {"site","no site"} for all categories.
3. one ore more bias maps are supplied. These are floating point (*i.e.* continuous values) type GIS maps in which every cell is assigned a value between "0" and "1". The values specify the degree to which it is believed that observed differences are biased towards the "no site" hypothesis. 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".
4. one or more attributes of the site vector point map are specified to represent the degree to which these points (sites) are biased towards the "site" hypothesis. Again, these attributes must be floating point values in the range 0 to 1. Calculations are similar to case 3: the more such biased sites are present on a certain category of an evidence map, the more mass will be subtracted from the "site" hypothesis and shifted to the uncertainty hypothesis.

In summary, a high amount of basic probability mass is shifted to the uncertainty hypothesis for map 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. Dempster's Rule of Combination allows us to combine any number of evidences and their belief mass distributions, including those parts assigned to individual

[93] To be precise, the complete set of hypotheses for this FoD also includes the empty set {null}, which is of mere mathematical necessity and will be of no further concern here.

uncertainty hypotheses. This provides an elegant and flexible framework for representing uncertainty in the site and landscape data, making it possible to create decision support maps based on the best available knowledge rather than on idealized models.

The following Dempster-Shafer output options constitute a complete set of tools for creating, exploring and interpreting predictive models, and for using them to make decisions in the face of incomplete and uncertain information:

- *belief(A)* is the total belief in hypothesis A. It tells us how much of the evidence speaks for A. This is the most basic DST function.
- *plausibility(A)* is the theoretical, maximum achievable belief in A. From a different point of view, it tells us how little evidence speaks against A (Ejstrud 2005).
- the *belief interval* represents the degree of uncertainty, measured as the difference between current belief and maximum achievable belief. It is defined as *plausibility(A)-belief(A)*. Areas with high belief intervals represent poorly studied or understood regions where additional/better information could improve model results (see Ejstrud 2003).
- a *weight of conflict* larger than zero indicates that evidences from different sources disagree with each other. A high weight of conflict might indicate a serious flaw in the model design or disagreement of evidences supplied by different data sources.

In the next section, we will demonstrate the use of these tools on the Borkeld case study.

9.3.2 A dempster-shafer predictive model for the borkeld study area

We can now proceed to examine the recreated Borkeld environmental and archaeological data set, perform some spatial data editing on it, and run the predictive modelling calculations. We will take as guiding principle that it is always better to improve these input data than to try to produce optimal weighting schemes for the locational factors: errors will be easier to control, results easier to interpret and model quality will automatically benefit from any future updates of the input data sets[94]. Therefore, the predictive model will be built only from basic measurements (*e.g.* elevation, hydrology) or from maps derived from those basic data using a formalized, well-known procedure (*e.g.* slope, aspect, visibility; Skidmore 1989). Since many maps in the Borkeld data set were derived from the basic geomorphology and soil maps in a non-standard way, using some form of subjective weighted overlays and classifications, they are highly correlated with the basic maps and would introduce unwanted overweight of certain variables if we were to use them as well. They are therefore excluded from this study.

Data used in the model

The archaeological information used in this study consists of three different sets of settlement site data representing our state of knowledge in 1989, 1991 and 2005. Of the full data set, 84 are Palaeolithic and Mesolithic sites, 59 Late Bronze Age to Early Medieval sites, and 164 Late Medieval sites (figure 9.11). One possible strategy would be to use all sites known in 1991 to build a predictive model, then test its performance with the additional sites known in 2005. However, there is a risk (especially in view of the small overall sample sizes) that the test data set itself will be biased due to the influence of specific forms of survey and localized developments. We therefore prefer to draw a random test sample of 33% from the total number of sites for each chronological phase, and set these aside for testing model performance later on (section 9.3.6). This leaves 57, 40 and 112 sites respectively for the actual modelling (table 9.5). Here, we will use only the Palaeolithic and Mesolithic sites to demonstrate how the DST model works.

[94] We will demonstrate these benefits in sections 9.3.4 and 9.3.5, when we look at how to add new information.

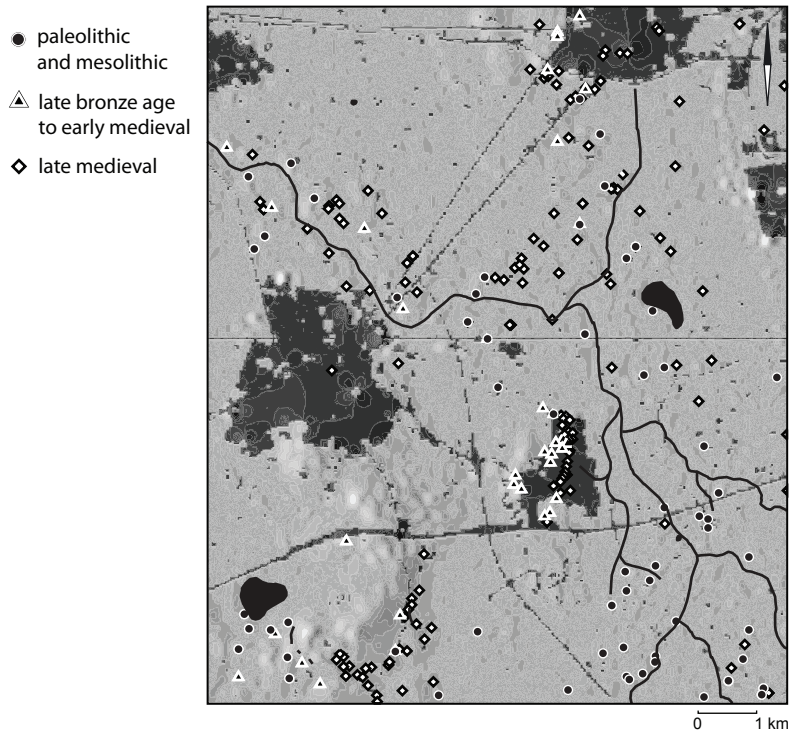

Figure 9.11 Overview of the Borkeld study area with built up areas (black), principal lakes and rivers, and all archaeological sites known in 2005.

no.	Type	total size	modelling size	testing size
1	Palaeolithic to Mesolithic	84	57	27
2	Late Bronze Age to Early Medieval Period	59	40	19
3	Late Medieval Period	164	112	52

Table 9.5 Summary of archaeological settlement site data sets used for the Borkeld DST case study.

The Borkeld data set also includes a lot of high-quality environmental data. A quick inspection of site distributions revealed the following sources of evidence to be of some relevance for site locations: height, slope, aspect, soil type, geomorphology, groundwater level, and hydrology.

Height, slope and aspect

A digital elevation model (DEM) was provided from centimetre precision LIDAR measurements. Unfortunately, the LIDAR sensors have also picked up all distortions caused by modern urban development and had to be cleaned before it could be used (figure 9.12; see comments below). The final "height" source of evidence was derived from the original (cleaned) 7 to 37 meter range by classifying the elevation measurements into 15 classes representing 2 meter steps.

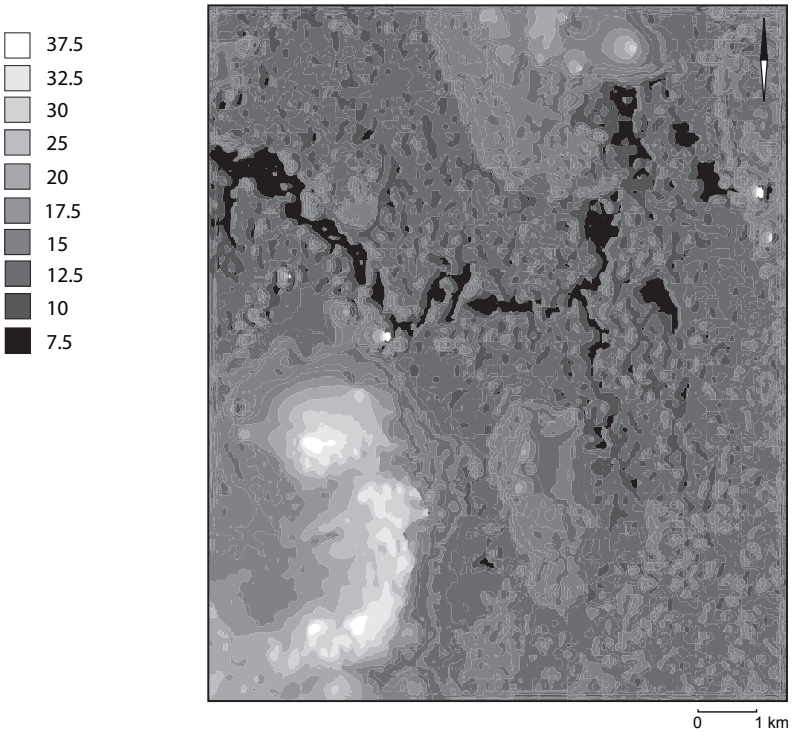

Figure 9.12 Cleaned digital elevation model. The height is in meters above sea level (a.s.l.).

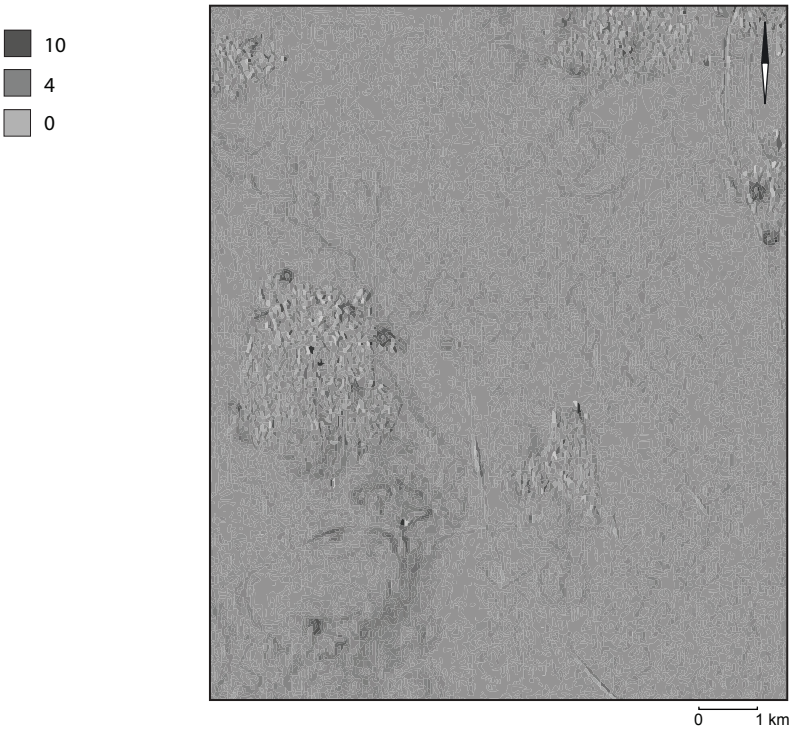

Figure 9.13 Slope steepness measure (in degrees), derived from cleaned elevation data.

A measure for the steepness of slopes can be derived directly from the cleaned DEM data (figure 9.13; see Horn 1981). Even in the relatively flat Borkeld landscape, a difference of a few degrees can have decisive impact on soil properties, such as susceptibility to erosion. Thus, a classified "slope" map using 0.5 degree steps was included in the model.

Aspect, *i.e.* the direction in which a slope is facing, can also easily be derived from the cleaned DEM using a standard GIS algorithm (again, see Horn 1981 for mathematical details). The main problem with the aspect variable is that it is circular, with 1 degree of bearing lying adjacent to 360 degrees and '0' representing totally flat areas (figure 9.14). The data was therefore reclassified into eight classes representing approximate N, NE, E, SE, S, SW, W, NW bearings in 45 degree steps.

Soil type and Geomorphology

The Borkeld data set contains a soil mapping with 11 classes (figure 9.15) and a very detailed map of geomorphology with 47 classes. Both were used in unaltered form in this case study.

Groundwater classes and Hydrology

Information about ground water depths and soil permeability was combined to create a map with three classes representing "very dry", "dry" and "wetter" soils (figure 9.16).

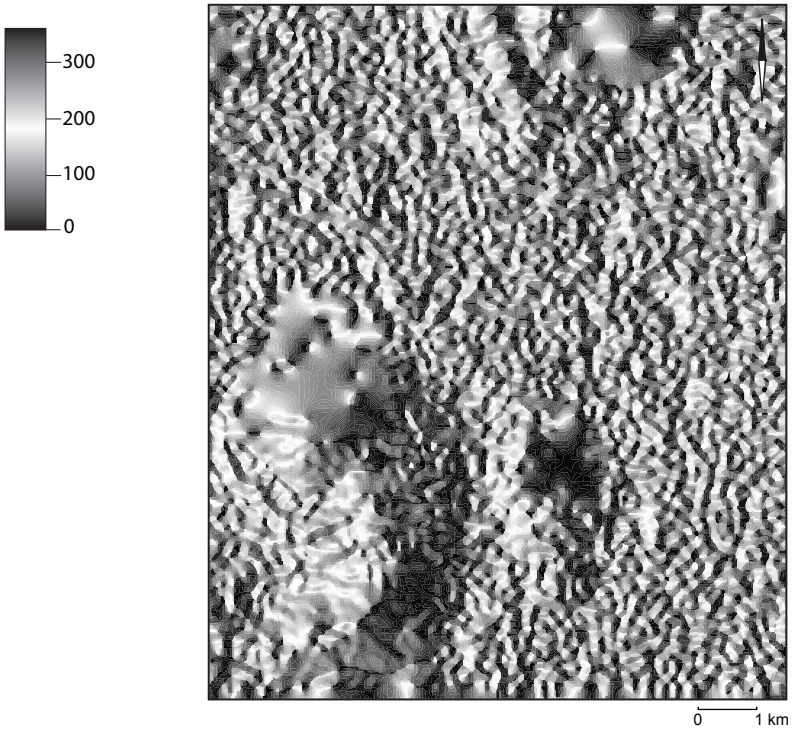

Figure 9.14 Aspect (bearing in degrees from East) derived from cleaned elevation data.

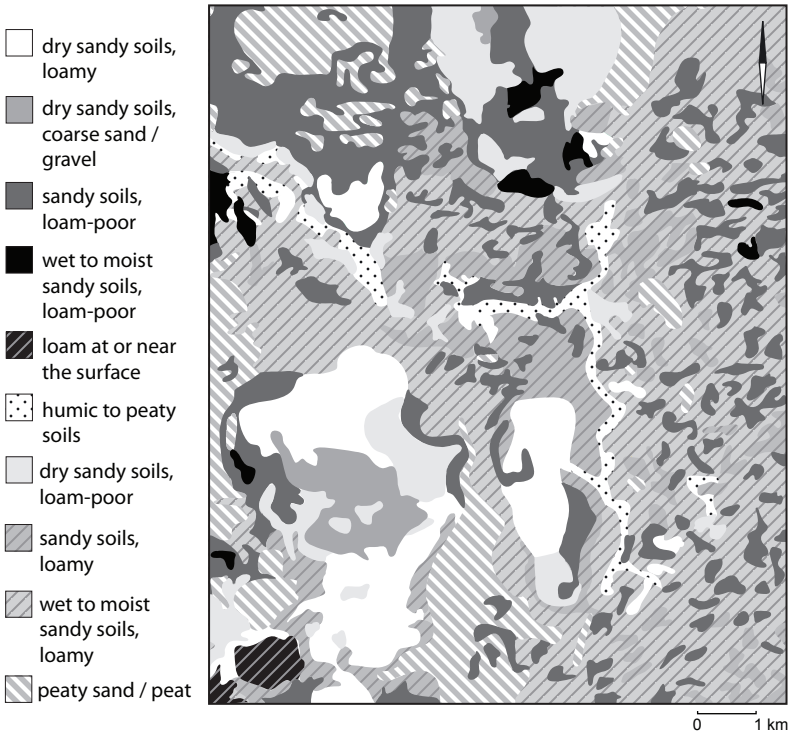

Figure 9.15 Soil type classes in the Borkeld study area.

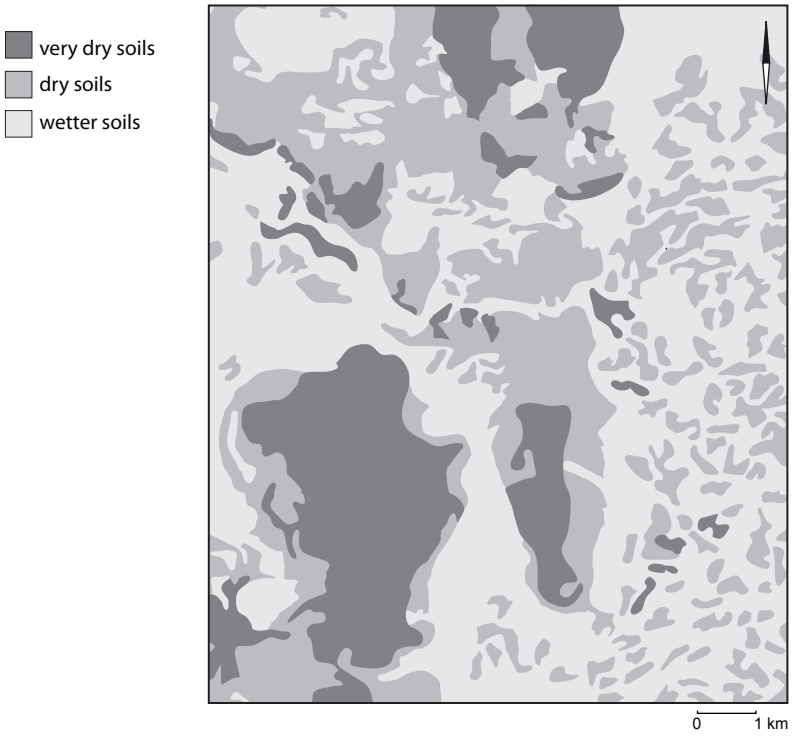

Figure 9.16 Groundwater classes in the Borkeld study area.

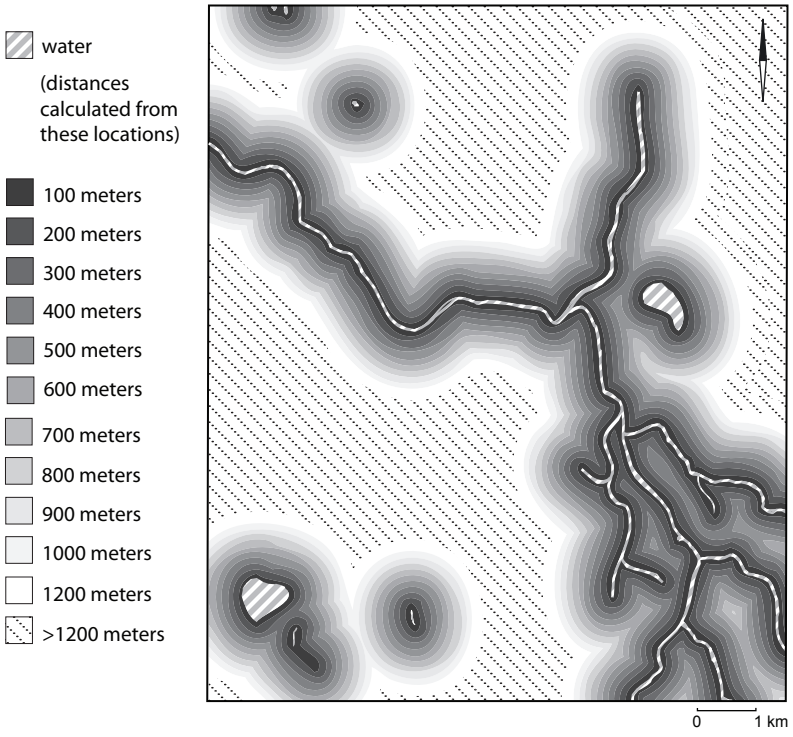

Figure 9.17 Distances from lakes and rivers in 100 m steps.

Proximity to open water for transportation, defence, nutrition and numerous other reasons is conventionally considered an important factor in locational behaviour. This case study uses the original Borkeld buffered hydrology map that has 100-m distance buffers around lakes and rivers (figure 9.17).

Altogether, this gives us seven environmental variables to capture a significant part of the locational behaviour reflected in the site data sets (table 9.6).

no.	name	data range	no. of classes	description
1	height	ca. 7.0 to 37.0 m	15 (2.0 m)	absolute elevation
2	slope	0.0° to 9.0°	18 (0.5°)	measure of slope steepness
3	aspect	1° to 360° + "0"	8 (45°)	direction of slope
4	soil	1 to 11	11	soil type classes
5	geomorphology	1 to 47	47	geomorphological classes
6	groundwater	1 to 4	4	groundwater depth classes
7	hydrology	1 to 17	17	distance classes (100 m) from open water
(8)	visibility	0.0 to 1.0	8	relative visibility from each location

Table 9.6 Summary of maps used as sources of evidence in the Borkeld DST case study. The last evidence,'visibility', will be introduced in section 9.3.5.

9.3.3 Interpretation of the results

The DST modelling software we used (a module especially written for GRASS GIS) includes all necessary tools to calculate Basic Probability Numbers for the data presented in the previous section, to combine evidences and to explore the results. A calculation of the four most important DST functions (Belief, Plausibility, Belief Interval, and Weight of Conflict) for our 57 Palaeolithic and Mesolithic sites gives four different maps for the 'site' and 'no site' hypotheses. We will here concentrate on the meaning of the maps for the 'site' hypothesis, keeping in mind that the belief outcomes for 'site' are not necessarily the inverse of 'no site', depending on how much uncertainty there is in the data.

Belief

The belief maps for 'site' and 'no site' constitute the basic predictive model information. They use a colour scheme going from black (lowest value) to white (highest value). The white patches thus represent areas where we would expect to find the most (figure 9.18) or the least (figure 9.19) buried archaeological sites, respectively. The belief maps for 'site' and 'no site' depicted here are not necessarily each other's exact complement, because some of the belief mass may still be assigned to the uncertainty hypothesis. Belief values, like all DST measures, are normalized to range from 0 to 1.

Plausibility and Belief Interval

As was stated before, plausibility is a measure of the maximum belief in a hypothesis that could be achieved if none of the uncommitted belief mass spoke against it. The belief interval represents the difference between belief and plausibility. These measures are of rather theoretical interest and will not be used for decision support in our scenario. However, the issue of uncommitted belief (uncertainty) is obviously relevant and will be discussed further in the next section.

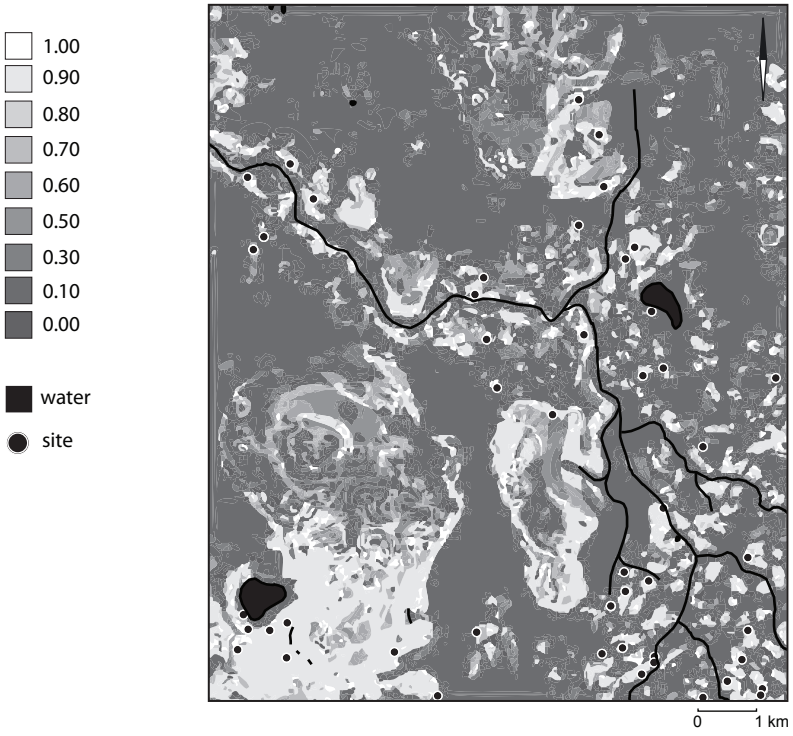

Figure 9.18 Map of belief in the hypothesis 'site' for Palaeolithic and Mesolithic sites. Principal lakes and rivers as well as positions of sites used in building the model are indicated.

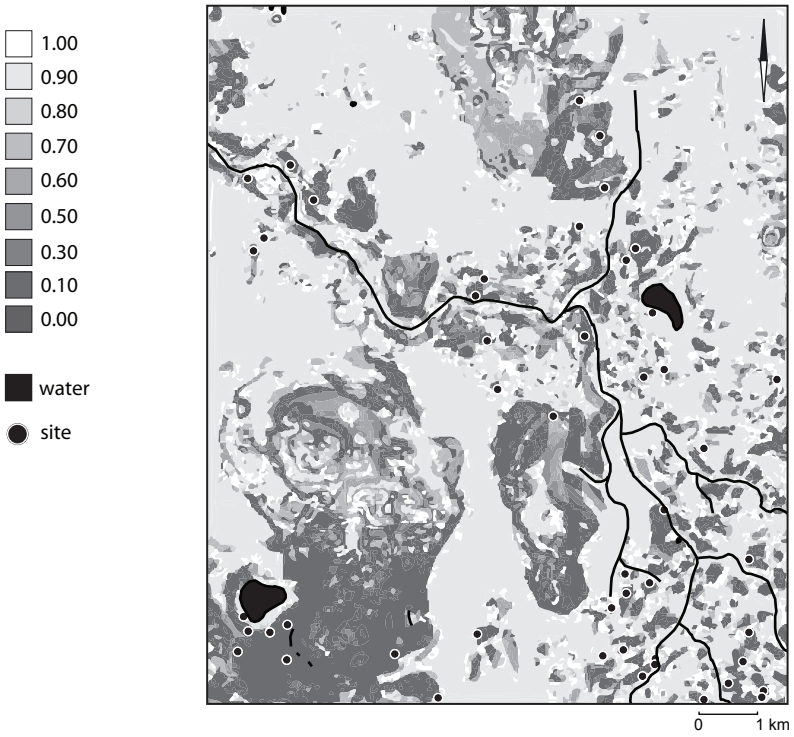

Figure 9.19 Map of belief in the hypothesis 'no site' for Palaeolithic and Mesolithic sites. Principal lakes and rivers as well as positions of sites used in building the model are indicated.

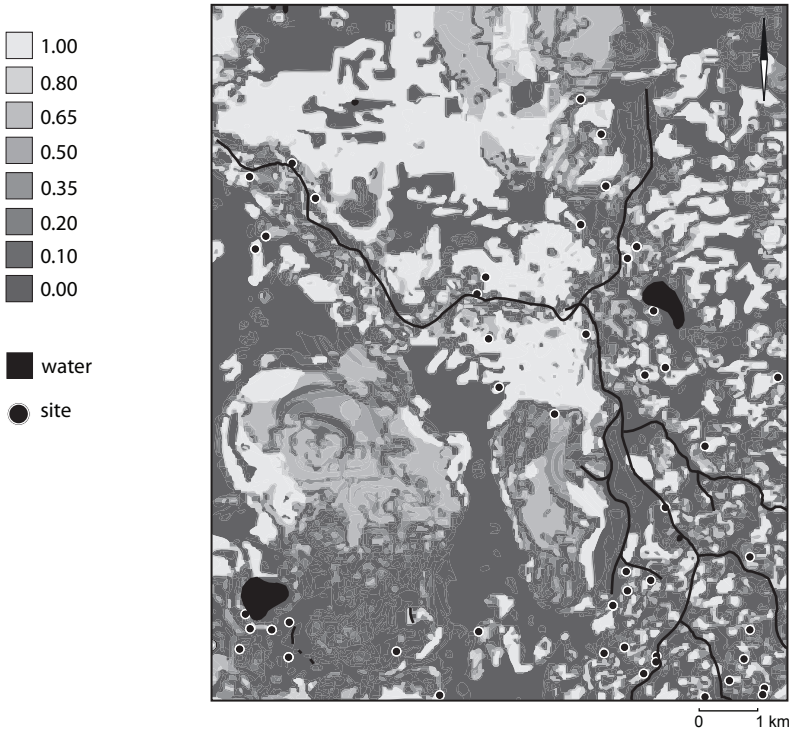

Figure 9.20 Map of the weight of conflict for all evidences for Palaeolithic and Mesolithic sites. Principal lakes and rivers as well as positions of sites used in building the model are indicated.

Weight of Conflict

The Borkeld model for the Palaeolithic and Mesolithic sites shows considerable weight of conflict in many areas, indicating that the evidence used is contradictory (figure 9.20). This may be due to the fact that the model would benefit from better knowledge about the relationships between the known sites and the environmental data; we will see later that this is indeed the case, and how the weight of conflict can be reduced for this specific set of sites. The contradictions may also be caused by the nature of the archaeological data set: it does not differentiate between sites of different periods and characters. By comparison, the weight of conflict map for the Late Bronze Age to Early Medieval sites (figure 9.21) seems to indicate that the selected sources of evidence are more appropriate for those sites.

To sum up, DST provides us with seven maps to explore the two basic hypotheses:

- *belief('site')*, *plausibility('site')*, *belief('no site')* and *plausibility('no site')* constitute the basic information body for decision support in impact mitigation scenarios.
- *belief interval('site')*, *Belief interval('no site')* and *weight of conflict* provide measures of uncertainty and contradiction and can be used to identify areas where caution is appropriate.

The information in these maps can be used to assess model performance; this is explored in section 9.3.6. But first we will have a closer look at how DST handles uncertainty and the addition of new evidence.

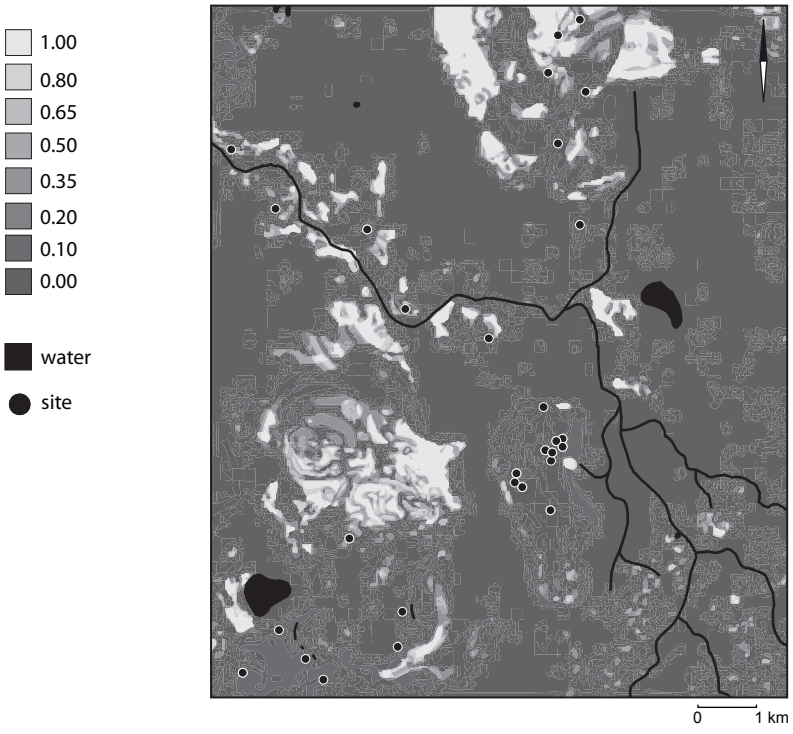

Figure 9.21 Map of the weight of conflict for all evidences for Late Bronze Age to Early Medieval sites. Principal lakes and rivers as well as positions of sites used in building the model are indicated

9.3.4 Extending the basic model: adding more uncertainty

At this point, it is time to explore the DST handling of uncertainty in more detail and to see how it can be used in practical management of cultural resources (CRM). Let us consider a basic CRM decision scenario: we are looking for an optimal location for a new, destructive development. CRM managers are required to point out areas of low archaeological value. As was argued before, the 'no site' belief map should be the basic tool in such a case. However, hard-earned experience justifies the assumption that this map is unreliable in its basic form, because the lack site information in hard-to-investigate parts of the study area may easily lead to an overrepresentation of 'no site' locations.

To quantify these doubts about the quality of our archaeological evidence, we will introduce some additional uncertainty related to land use categories in the study area (table 9.7 and figure 9.22). By asking each of our regional experts to rate the effect of each land use type on the discovery of archaeological sites, we obtain a bias number that, once normalized to a value between 0 (not biased towards 'no site') and 1 (highest bias towards 'no site'), can be mapped and used to alter our belief in the 'no site' hypothesis.

It should be noted that the bias quantification demonstrated here is merely for illustrative purposes. Our intention is simply to show the principal method of introducing uncertainty into a Dempster-Shafer predictive model. Thus, the bias values for different land use classes as listed in table 9.7 remain subjective and poorly founded.

	class	simplified class	expert 1	expert 2	expert 3	bias
0	built up	urban	0.4	0.5	0.6	0.4
1	grassland	pasture	0.8	0.6	0.7	0.7
2	deciduous woodland	woodland	0.7	0.8	0.8	0.7
3	coniferous woodland	woodland	0.7	0.8	0.8	0.7
4	maize, grain	arable	0	0	0	0.1
5	water	water	0.9	0.9	0.9	0.9
6	potatoes, beets	arable	0	0	0	0.2
7	other crop	arable	0	0	0	0.2
8	heather	moor	0.6	0.8	0.8	0.5
9	bare soil	-	-	-	-	0.0

Table 9.7 Land use classes of the Borkeld area, with experts' opinions regarding 'no site' bias, and the bias weights used here to alter the DST model.

A reclassification of the land use map yields the spatial distribution of 'no site' bias. Comparing the old and new 'no site' maps side-by-side (figure 9.23; compare with figure 9.19) shows how this results in a much lower confidence, but it also shows, more realistically, where additional field research should be directed to reduce the (commercial and scientific) risks. In fact, a summary of the 'belief interval' values for this model shows that up to one third of the evidence information remains on the uncertainty hypothesis, even though seven considerably detailed sources of information were used. In the next section we will show that this uncertainty can be reduced by introducing new relevant evidence, and excluding evidences that turn out to have little relevance for the particular set of sites being modelled.

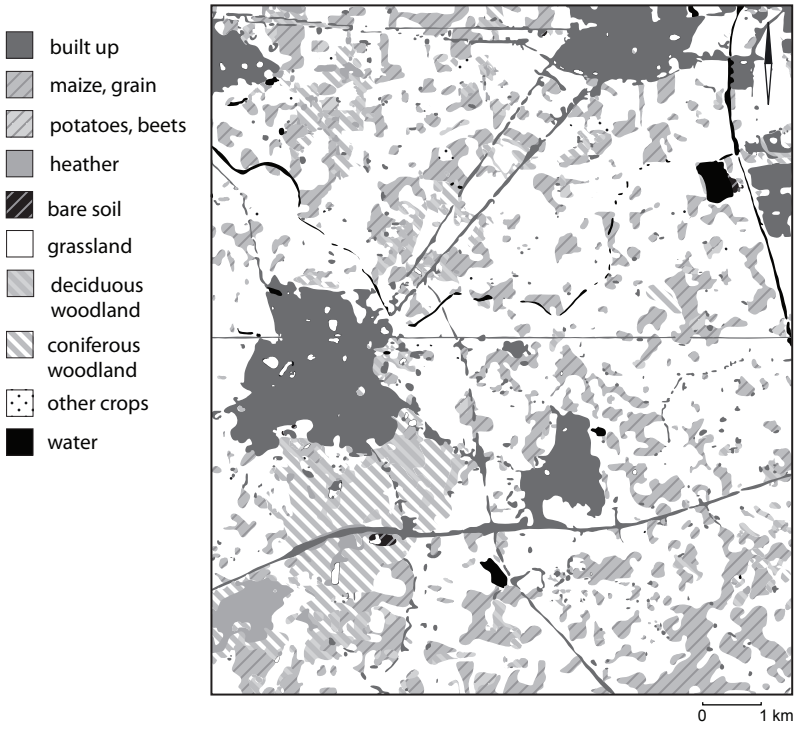

Figure 9.22 Categories of land use in the Borkeld study area.

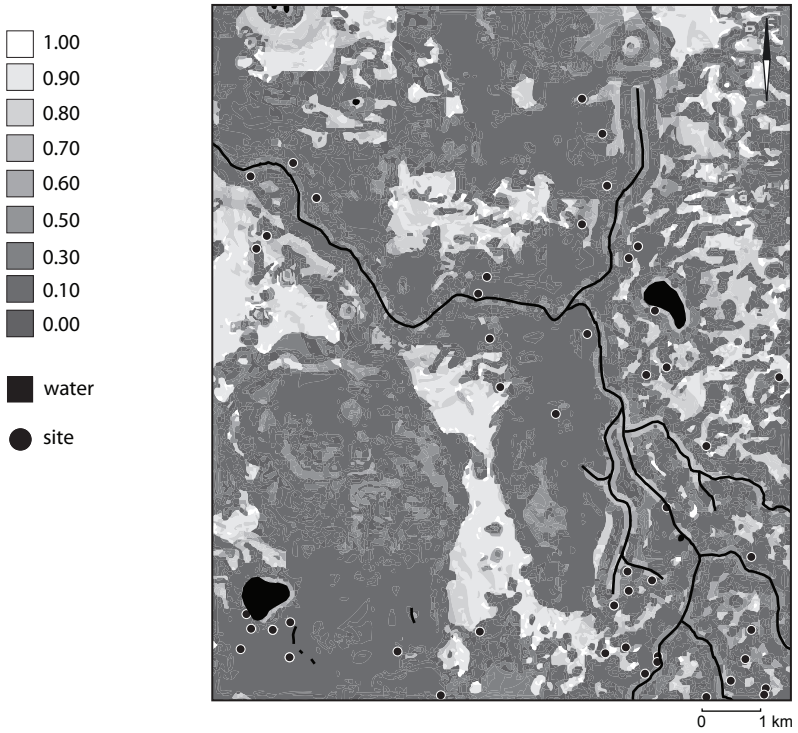

Figure 9.23 Map of belief in the hypothesis 'no site' for Palaeolithic and Mesolithic sites, as weakened by bias derived from land use classes. Principal lakes and rivers as well as positions of sites used in building the model are indicated. Compare with figure 9.19.

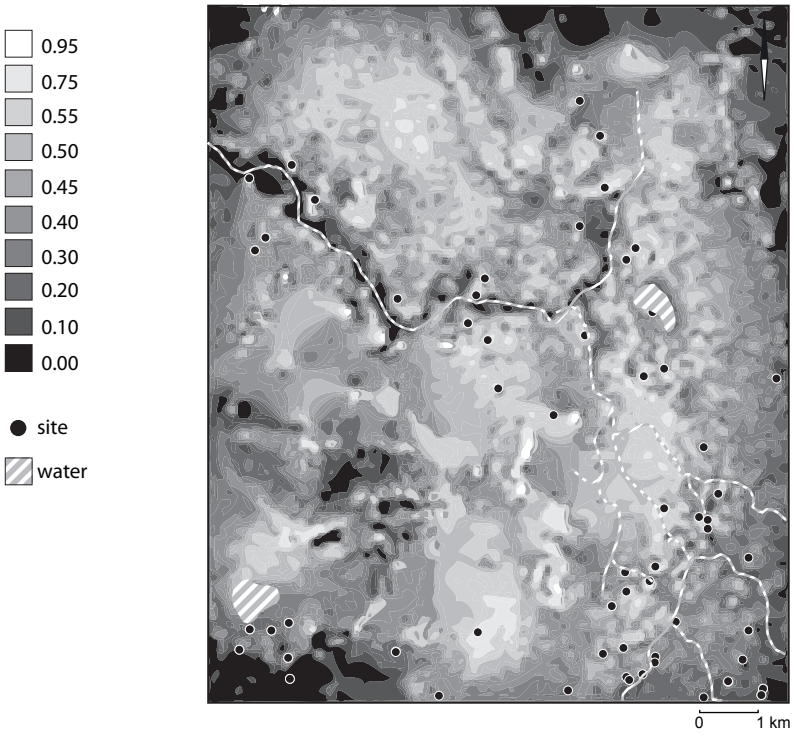

Figure 9.24 Normalized visibility, derived from a cumulative viewshed analysis for the Borkeld test area. The positions of all Palaeolithic and Mesolithic sites known in 2005 are indicated.

9.3.5 Updating the basic model: optimizing evidences

Earlier, it was claimed that a DST model can be easily improved by updating the information fed into it. By updating, we mean providing better, more accurate information or adding new, relevant sources of evidence. Here we will demonstrate the updating process for our model of the Palaeolithic and Mesolithic sites by adding visibility (viewshed) information as another source of evidence, and excluding the evidences 'aspect' and 'groundwater depth' which we suspect have little locational relevance for this specific group of sites.

Regarding the Palaeolithic and Mesolithic sites, both theory and an inspection of the available data suggests that a fair portion of them prefers "observer point" locations with a good view of the surroundings. We will therefore include visibility information, derived from a viewshed analysis of our elevation data, to improve the model results for this group of sites[95]. The raw visibility data are first normalized to a range of 0 to 1, representing lowest and highest visibility from a specified point (figure 9.24), and then broken up into eight categories of equal width to represent the 'visibility' piece of evidence in our DST model.

A model run with this updated information confirms the validity of these choices: model gain is improved from 0.68 to 0.73 as the evidence now includes more relevant information (figure 9.25, compare with figure 9.18; table 9.9, compare runs 1 and 7), and the weight of conflict map for the revised model (figure 9.26) shows that the sources of evidence are now more in agreement with each other than they were before the revision (figure 9.20).

[95] The visibility model used here was generated by a cumulative viewshed analysis in GRASS GIS: every cell in the raster model records the number of cells visible from that cell within a 2 km radius. The accuracy of the result suffers from the usual defects, such as edge effects, errors in the DEM etc. However, for our purposes it is sufficient to have a rough model of visibility.

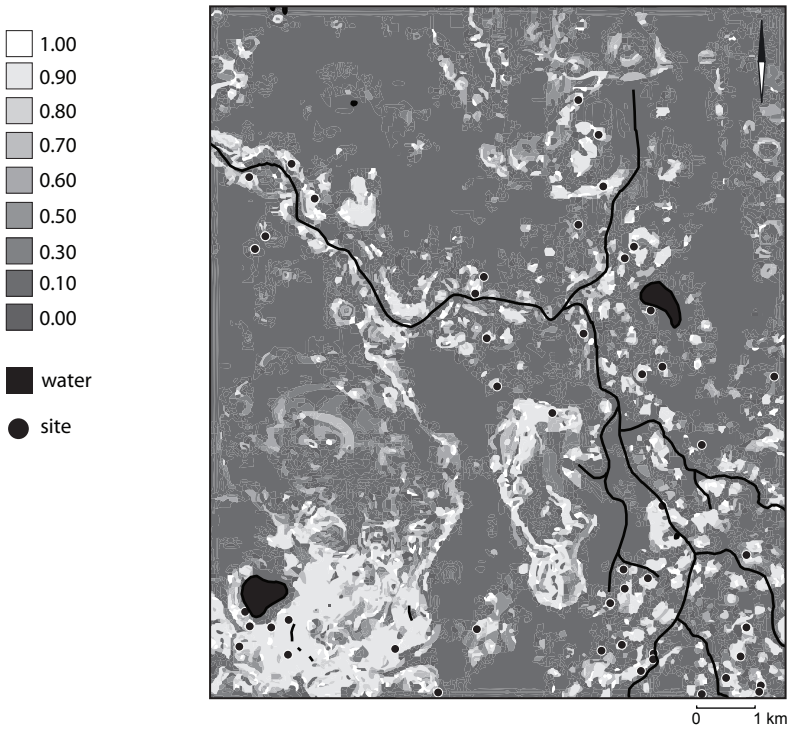

Figure 9.25 Belief map ('site') for the improved Palaeolithic and Mesolithic sites predictive model (compare with figure 9.18).

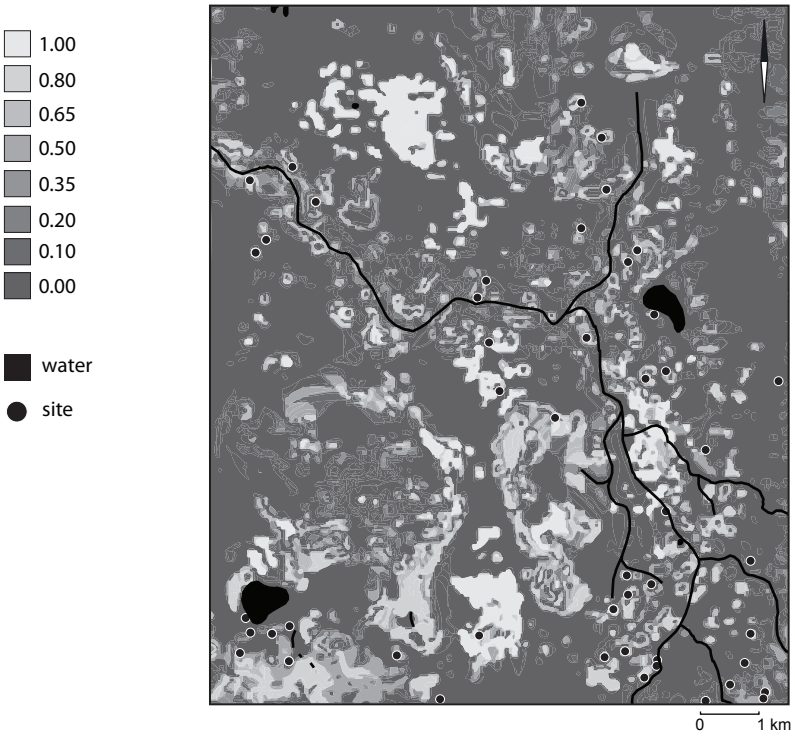

Figure 9.26 Weight of conflict map for all sources of evidence for the improved Palaeolithic and Mesolithic sites predictive model (compare with figure 9.20).

9.3.6 Testing the models

It is now time to assess the performance of our DST predictive models for the Borkeld area. For demonstration purposes, we will concentrate on the belief maps for the 'site' hypothesis, and classify the belief values into just three decision support ranges. In what follows, the "low" range represents belief values from 0.0 to 0.33, "medium" from 0.34 to 0.66 and "high" from 0.67 to 1.0.

From a CRM point of view, sites located in the "high" belief range are at low risk, because the belief map correctly indicates the high archaeological value of their geographic setting. Of course, the "high" belief range itself should not take up too large a share of the model area; society would not accept the costs and limitations imposed by such an overprotection of the archaeological heritage. A good predictive model must therefore be able to place as large an amount of sites in as small an area of "high" belief as possible[96]. This ability is expressed by Kvamme's gain statistic, defined as *1 - %area / %sites*, and should be as close to 1 as possible. For example, a model which classifies 50% of the total area as "high" belief and captures 100% of the sites within that area will only achieve a gain of 0.5.

Likewise, from a CRM point of view, sites located in "low" belief areas represent failures of the predictive model. In a real-world scenario, these site would be under a severe threat of being destroyed or under-managed. The number of sites in the "low" belief areas is therefore a direct measure of the model's practical reliability. Finally, the "medium" belief range is hardest to interpret and use in a CRM context; for this reason our models should attempt to keep the share of the "medium" range as small as possible.

It would be desirable, of course, to have more detailed, quantitative measures for the different aspects of model performance. Currently, Kvamme's gain statistic is the only one widely used in predictive modelling applications, although a number of additional ones have been published recently (Kvamme 2006).

Table 9.8 shows the performance results for the modelling sets of sites for the three periods. We can use this to estimate the model's maximum expected performance given specific input data. For each site data set, a number of statistics are shown that relate the site datasets to the predicted areas of "low", "medium" and "high", respectively. The columns 'sites' and '%sites' show the number and percentage of sites located in each belief range. These should be as low as possible for "low" belief areas and as high as possible for "high" belief areas. The columns 'cells' and '%cells' show how much of the total study area is taken up by each of the belief ranges; the areas of "medium" and "high" belief should be as small as possible. '%sites' and '%cells' are used to calculate Kvamme's gain, and in this column the number of greatest relevance is the one for the "high" belief areas, which should be as close to 1 as possible. When we next test the performance of our model using the test data sets set aside at the start, we may expect best results for the Late Bronze Age to Medieval Period, somewhat worse for the Palaeolithic to Mesolithic, and worst of all for the Late Medieval Period.

How well do our models handle the 'new' site information from our test data sets? Table 9.9, runs 1 to 3, shows that our models perform only very slightly worse for the test data than they did for the original data sets. The loss in "high" belief gain is only 0.01 to 0.02, indicating that our models do a good job generalizing from the original data sets. This represents a significant improvement over the original 1990 predictive model (Ankum and Groenewoudt 1990), which degraded rapidly when checked against the additional site data that had become available by 2005 (*e.g.*, almost half the Palaeolithic and Mesolithic sites were missed).

The overall performance of our model is worst for the Late Medieval sites, with a gain of only 0.56. Apparently, the sources of evidence chosen do not adequately reflect the locational behaviour in our Late Medieval sites data set, or the data set contains sites of different types and locational properties. This problem was already identified in the first Borkeld case study (Ankum and Groenewoudt 1990), and serves as a reminder of why it is important to split site data into temporal and functional classes before model building.

[96] see this volume chapter 8

cat.	description	sites	% sites	cells	% cells	gain
			Palaeolithic to Mesolithic			
0	low belief("site")	10	17.86	131210	69.05	-0.74
1	medium belief("site")	3	5.36	13977	7.36	-0.27
2	high belief("site")	43	76.79	44829	23.59	0.69
			Late Bronze Age to Early Medieval Period			
0	low belief("site")	5	12.50	139033	73.17	-0.83
1	medium belief("site")	2	5.00	10642	5.60	-0.11
2	high belief("site")	33	82.50	40341	21.23	0.74
			Late Medieval Period			
0	low belief("site")	26	24.53	124420	65.48	-0.63
1	medium belief("site")	8	7.55	10884	5.73	0.24
2	high belief("site")	72	67.92	54712	28.79	0.58

Table 9.8 Performance of predictive models with the original site data used to build the models.

Runs 4 to 6 of table 9.9 show what happens to model performance when we introduce the 'no site' bias discussed in section 9.3.4: the increased uncertainty in our model weakens our ability to support the 'site' hypothesis to such a degree that the "high" gain drops to values of 0.24 to 0.45 – too low to be useful in decision making. However, there is also almost no risk of missing a site when making a "low archaeological value" decision, as the number of sites in those areas is now insignificantly small.

run	test data set	sites	"high" area	"high" gain	misses
1	Palaeolithic to Mesolithic	27	23.59%	0.68	5
2	Late Bronze Age to Early Medieval Period	19	21.23%	0.73	4
3	Late Medieval Period	52	28.79%	0.56	14
4	Palaeolithic to Mesolithic, biased	27	70.17%	0.24	1
5	Late Bronze Age to Early Medieval Period, biased	19	51.64%	0.45	1
6	Late Medieval Period, biased	52	70.29%	0.27	2
7	Palaeolithic to Mesolithic, visibility added	27	18.72%	0.73	5

Table 9.9 Performance of predictive models as indicated by validation with the test data set.

Finally, run 7 of table 9.9 shows the effect of improving the evidence information on performance. The addition of the 'visibility' evidence and the exclusion of 'aspect' and 'groundwater depth' have resulted in an 0.07 point increase of the "high" gain as compared with run 1.

9.4 Discussion

It was the purpose of this case study to demonstrate two alternative approaches to the incorporation of uncertainty and expert judgment into formal predictive archaeological models. Out of a number of possibilities, we chose to pursue Bayesian analysis and Dempster-Shafer theory based on our earlier study of relevant published research. We enlisted the help of three archaeological experts to provide us with expert judgments about location fcators and bias factors for the case study area of the Borkeld (Rijssen-Wierden), scene of the first published archaeological predictive model in the Netherlands.

In view of the technical complexity of our demonstration models, we will here discuss not just how predictive models incorporating uncertainty may be of use to actors in the arena of archaeological heritage management, but also the problems we have faced in obtaining and evaluating expert opinions and in deciding whether any particular approach or model is 'better' than another.

Models for decision support

The models resulting from both of the approaches described here can be used as *decision support tools*, for example allowing those who are responsible for setting briefs to be more specific than has been possible until now: 'look *here* for Bronze Age habitation sites', or 'we do not have enough information yet about *this* area yet; conduct a preliminary field assessment to gather these data'. In a project covering a sufficiently large area, statements of the probability of finding sites, or particular types of site, can be used directly to estimate the likely *number* of sites and to prepare budgets. However, on smaller development projects probability statements are less likely to be of direct use, and it is more likely that some sort of preliminary investigation ('assessment" in the English Heritage jargon) will be needed. In this case the methodology developed by Orton and Nicholson (Orton 2000a; 2000b; 2003) can use a probability statement to design an assessment to ensure with a specific confidence level that, if no archaeological remains are found, there are none there. A probabilistic statement thus has the desirable property of being able to directly feed into an algorithm for designing fieldwork strategies for assessment or mitigation.

How 'uncertainty' approaches are going to be used, depends on the *purpose* of the model. For example, if the model is to be used to define zones of high archaeological expectation for municipal zoning schemes, then a simple division into 'low' and 'high' zones is required. If the subsidiary purpose then is to reduce as much as possible the area of the 'high' zone, then the model should aim at maximizing *gain* for the 'high' zone; if on the other hand the subsidiary purpose is to avoid, as much as possible, the uninvestigated destruction of sites, then it should aim to minimize *gain* for the 'low' zone. For predictive models to become true decision support tools, the 'decider' at local or regional government level must be able to set this kind of preference.

If, furthermore, predictive models are produced for each of a number of site type/period combinations (*e.g.*, 'Mesolithic hunting camps'), then the 'decider' can use his own predefined priorities to select some of these (and ignore others). Since each of the input models would incorporate the best available data and expertise, and each also provides absolute probabilities/risks as output, it should be possible to logically combine them (*e.g.*, using binary OR) into a final model.

Another important requirement of predictive models as decision support tools is that they should be easily *updatable* with new evidence and expertise. Additional archaeological observations and minor changes to evidences (*e.g.* a finer scale soil map becomes available) in both Bayesian and DST approaches are simply used to re-calculate and update the model, but adding new evidences or substantially changing existing ones would also require new expertise to be generated. A workable solution would require that each new evidence is 'packaged' with metadata containing the appropriate expertise.

Quantifying expertise

For the Bayesian model, getting the required information from the experts can be somewhat of a struggle. The experts were asked to generalize from particulars, and to quantify aspects that they are used to thinking about in qualitative mode. However, we want to stress that the use of multiple experts, by disagreeing on some of these quantitative weights (*e.g.*, the *distance to water* weight), provide us with a relatively objective measure of uncertainty. So, should we conclude that the more experts we can get, the better our models will be? No, because this introduces the question of the *relative expertise* of the experts: given that different experts have different types and degrees of expertise with relevance to the model being built, we would need to rate them, or their opinions, or both. We also need to specify how we rate expertise against actual data: how much data is required before we disbelieve the expert? The 'data equivalent' calculated in section 9.2 for our Bayesian analysis provides a first peek into what we may call the field of *meta-expertise*, but clearly this will need to be more thoroughly studied.

Experts' opinions about the spatial patterning of archaeological remains does not simply fall from the sky – they are based on both theoretical considerations and personal or collective observations. To the extent that expert opinion is not independent of actual observations, its 'weight' should be lowered. Bias factors such as land use and land cover, discovery mode, and geological history have such a great impact on the patterning of known archaeological sites that researchers must put more effort into understanding and quantifying these factors. We recommend that the potential of non-site observations be studied in the context of linear developments.

Similarities and differences of DST and (Bayesian) probability models

Naturally, there are numerous similarities between DST and (Bayesian) probability theory, as both approaches try to provide an abstract framework for the same basic problem: reasoning using uncertain information. A comparison of the formulas for both Dempster's Rule of combination and Bayes' Theorem reveals their mathematical kinship; in fact, DST can be viewed as a generalization of Bayesian theory (Dempster 1968). Probabilistic models have the advantage that their output can be used straightforwardly in testing the model with new data (see also this volume chapter 8).

One significant practical difference between the two approaches is that Bayesian strictness is not *required* in a DST model: much simpler ways of quantifying evidence, such as ratings, have also been shown to work (*e.g.* Lalmas 1997). DST is geared towards practical applicability: making decisions under uncertain conditions with information that is hard to formalize as probabilities (see Kuipers, Moskowitz and Kassirer 1988 for a much-cited example). This is why the term *beliefs* was introduced to replace *probabilities* in a DST context. DST is also more *explicit* in its representation of uncertainty as a model input. It is very easy to introduce uncertainty in a controlled way into a DST model and the output will reflect this directly, thus providing more realistic decision support for real-world CRM scenarios.

Conversely, adopting the more strict Bayesian approach means that it is absolutely clear what the quantified results mean – be they predictions of site densities or maps of our degree of confidence in these predictions. But there are further advantages to the Bayesian approach that address other problems and questions raised in the Baseline Report (van Leusen *et al.* 2005). For example, hierarchical models can be constructed which allow the aggregation of differing but similar classes of samples in a single analysis without assuming that they are identical. This could allow models to be built which account for chronological and functional subsets of sites in the landscape, whilst achieving the sample sizes necessary to achieve useably small posterior confidence ranges.

Non-site bias

It was observed in the Baseline Report that the failure of heritage managers both in the Netherlands and internationally to systematically record 'non-site' observations deprived model-builders of an important source of evidence. Observed 'non-site' locations, *e.g.* fields that were surveyed under good visibility conditions and where no archaeology was observed, can under certain conditions be used to estimate the probability of a site

occurring in a particular landscape setting. However, observed low site density in one suitable zone cannot be extrapolated to the next suitable zone, because people only ever use part of the suitable zone. So, to be useful, nonsite observations must be *random* to be representative for the study area. Although archaeologists have not systematically reported nonsite observations in the past, thousands of commercial archaeological assessments reported under 'Malta' legislation over the last decade present a good potential source for them. If such observations were incorporated in the posterior probability estimates of a Bayesian model, the uncertainties in the posterior uncertainty map could be lowered. For DST models, nonsite observations would strengthen the belief in the 'no-site' hypothesis for a specific environmental setting, and would reduce the remaining uncertainty. Given that nonsite locations are in the vast majority, and random observations are therefore likely to be nonsite observations, a *single* nonsite observation can only reduce uncertainty by a very small amount. We therefore need to collect large numbers of nonsite observations.

Outlook

Archaeological Heritage Managers in the Netherlands must make do with a relatively weak set of tools when it comes to assessing the archaeological 'potential' of an area. We have demonstrated, however, that the development of a flexible decision support tool based on GIS functionality is feasible. It remains to be seen, however, whether the 'field' perceives the need for a thorough revision of current procedures and products. The publication by the State Service for Archaeology, the Cultural Landscape and Monuments, in early 2008, of the third 'generation' of the Indicative Map of Archaeological Values of the Netherlands (IKAW; Deeben 2008), does not point in this direction.

References

Ankum, L.A. and Groenewoudt, B.J. 1990. *De situering van archeologische vindplaatsen*. RAAP-rapport 42. Amsterdam: Stichting RAAP

Brandt, R.W. 1987. *De waardering van archeologische vindplaatsen in het landelijk gebied*. RAAP-notitie 5. Amsterdam: Stichting RAAP

Brandt, R.W. 1990. De archeologische potentiekaart: anticiperen op dekwetsbaarheid van het bodemarchief. In J.H.F. Bloemers, C.W. van Pelt and F.A. Perk (eds), *Cultuurhistorie en milieu in 2015: op weg naar eenlandschap zonder verleden? Symposium ter gelegenheid van het vijffjarig bestaan van de Stichting Regionaal Archologisch Archiverings Project (R.A.A.P.)*, 1 maart 1990, 58-63. Amsterdam: Stichting RAAP

Brandt, R.W., B.J. Groenewoudt and K.L. Kvamme 1992. An experiment in archaeological site location: modelling in the Netherlands using GIS techniques. *World Archaeology* 24, 268-282

Deeben, J.H.C. (ed.) 2008. *De Indicatieve Kaart van Archeologische Waarden, derde generatie*. Rapportage Archeologische Monumentenzorg 155. Amersfoort: RACM

Dempster, A. P. 1967. Upper and Lower Probabilities Induced by a Multi-valued Mapping, *Annals of Mathematical Statistics* 38, 325–339

Dempster, A. P. 1968. A Generalization of Bayesian Inference, *Journal of Royal Statistical Society*, Series B 38, 205—247

Ejstrud, B. 2003. Indicative Models in Landscape Management: Testing the Methods. In J. Kunow and J. Müller (eds), *Landschaftsarchäologie und geographische Informationssysteme: Prognosekarten, Besiedlungsdynamik und prähistorische Raumordnungen. 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, 119-134. Wünsdorf: Brandenburgisches Landesamt für Denkmalpflege und Archäologisches Landesmuseum

Ejstrud, B. 2005. Taphonomic Models: Using Dempster-Shafer theory to assess the quality of archaeological data and indicative models. In M. van Leusen and H. Kamermans (eds), *Predictive Modelling for Archaeological Heritage Management: A research agenda*. Nederlandse Archeologische Rapporten 29, 183-194. Amersfoort: Rijksdienst voor het Oudheidkundig Bodemonderzoek

Gibson, T.H. 2005. Off the Shelf: Modelling and management of historical resources. In M. van Leusen and H. Kamermans (eds), *Predictive Modelling for Archaeological Heritage Management: A research agenda*. Nederlandse Archeologische Rapporten 29, 205-223. Amersfoort: Rijksdienst voor het Oudheidkundig Bodemonderzoek

Horn, B. K. P. 1981. Hill Shading and Reflectance Map. *Proceedings of the IEEE* 69(1), 14-47

Kuipers, B., Moskowitz, A. J. and Kassirer, J. P. 1984. Critical Decisions under Uncertainty: Representation and Structure. *Cognitive Science* 12, 177-210

Kvamme, K. L. 2006. There and Back Again: Revisiting Archaeological Location Modeling, GIS and Archaeological Site Location Modeling. In M. Mehrer and K. Wescott (eds), *GIS and Archaeological Predictive Modeling*, 3-38. Boca Raton: CRC Press

Lalmas, M. 1997. Dempster-Shafer's Theory of Evidence Applied to Structured Documents: Modelling Uncertainty. In *Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information*, 110-118. New York: ACM

Leusen, P.M. van, J. Deeben, D. Hallewas, P. Zoetbrood, H. Kamermans and Ph. Verhagen 2005. A Baseline for Predictive Modelling in the Netherlands. In M. van Leusen and H. Kamermans (eds), *Predictive Modelling for Archaeological Heritage Management: A research agenda*. Nederlandse Archeologische Rapporten 29, 25-92. Amersfoort: Rijksdienst voor het Oudheidkundig Bodemonderzoek

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

Orton, C., 2000a. *Sampling in Archaeology*. Cambridge Manuals in Archaeology. Cambridge: Cambridge University Press

Orton, C., 2000b. A Bayesian approach to a problem of archaeological site evaluation. In K. Lockyear, T.J.T. Sly and V. Mihailescu-Birliba (eds), *CAA96. Computer Applications and Quantitative Methods in Archaeology*. BAR International Series 845, 1-7. Oxford: Archaeopress

Orton C. 2003. *The Fourth Umpire: Risk in archaeology* (Inaugural Lecture)
http://www.ucl.ac.uk/archaeology/special/orton-inaugural-2003/index.htm, accessed 9 May 2003

Rogerson, P.A. 2001. *Statistical Methods for Geography.* London: Sage Publications

Shafer, G. 1976. *A Mathematical Theory of Evidence*. Princeton: Princeton University Press

Skidmore A.K. 1989. A comparison of techniques for calculating gradient and aspect from gridded elevation data. *International Journal of Geographical Information Systems* 3, 323-334

Smets, P. 1994. What is Dempster-Shafer's model?, In R.R. Yager, J. Kacprzyk and M. Fedrizzi (eds), *Advances in Dempster-Shafer Theory of Evidence*, 5-34. New York: Wiley

Smets, P. 2005. Decision Making in the TBM: the Necessity of the Pignistic Transformation. *International Journal of Approximate Reasoning* 38, 133-147

Smets, P. and Kennes, R. 1994. The transferable belief model. *Artificial Intelligence* 66, 191-234

Soonius, C.M. and A. Ankum 1991a. *Archeologische meldingskaart en potentiekaart van de gemeente Ede: een samenvatting*. RAAP-notitie 36. Amsterdam: Stichting RAAP

Soonius, C.M. and L.A. Ankum 1991b. *Ede*. RAAP-rapport 49. Amsterdam: Stichting RAAP

Appendix - the mathematical framework of dempster-shafer theory

This is a very brief introduction to the mathematical framework of DST. The full background can be found in the original publication by Shafer (1976) and, perhaps more accessibly, in numerous papers by Smets and colleagues (*e.g.* Smets 1994; 2005; Smets and Kennes 1994).

First, we will consider an abstract, standard situation in scientific research: building a model from a number of hypotheses and variables. Starting from scratch, there are:

- the basic *hypotheses*. They cover all possible outcomes of the model calculations.
- a number of variables which are deemed to be of importance for the model.
- a method to quantify the degree of support that variables lend to specific hypotheses (probabilities, rankings, etc.).

For each hypothesis, it is now necessary to check to what degree the provided variables support or refute it and calculate the total degree of *belief* in that hypothesis. Note that we are not talking about the *probability* of a hypothesis being true, as that would imply using the more rigid mathematical framework of probability theory; the concept of belief in DST is significantly more flexible.

The mathematical framework for the outlined procedure is provided by DST. At this point, we need to make some terminological definitions:

- the set of hypotheses $H=h_{1..n}$, which represent *all* possible outcomes, is called *Frame of Discernment* (FoD).
- a variable with relevance to the FoD is a *source of evidence*. The entirety of sources of evidence is called *body of evidence*. A variable value is transformed into an *evidence* by calculating a *Basic Probability Number (BPN) for it (this is also referred to as a basic probability assignment)*.
- a BPN is the basic quantification of evidence in the DST. It consists of a value m_n in the range "0" to "1" for each hypothesis in the FoD. The restriction is that $m_{1..n}$ must sum to "1", *i.e.* the entire basic probability mass must be distributed over the given FoD.

BPNs can be assigned to a singleton hypothesis in H as well as to subsets of it. What this means is that DST has the ability to represent *uncertainty* as subsets of H. Thus, if two hypotheses h_1="a" and h_2="b" are supplied, there will always also exist a set of hypotheses $\{h_1,h_2\}$ for the belief that both could be true ("a OR b"). In fact, this is the defining property of DST as a theory of uncertainty; probability theory, as used by the Bayesian model, does not have this explicit representation of uncertainty. What this means for the case of archaeological predictive models will be discussed later.

Any number of sources of evidence can be combined using *Dempster's Rule of Combination*:

$$m(A) = m_1 \oplus m_2 = \frac{\sum_{B \cap C = A} m_1(B) m_2(C)}{\sum_{B \cap C \neq 0} m_1(B) m_2(C)}$$

Dempster's Rule computes a measure of agreement between two sources of evidence for various hypotheses in the FoD (in the formula above: A,B,C). It "focuses only on the hypotheses which both sources support" (Lalmas 1997). The denominator normalizes the result to ensure that it stays in the range [0;1]. Any number of evidences can be combined by running the calculation recursively. This calculation can be computationally very expensive for a large number of hypotheses. From the result, a number of useful DST measures can be calculated. The most important ones are *belief* and *plausibility*.

The *belief function Bel(H)* computes the total belief in a hypothesis *A*. The total belief in *A* is the belief mass of *A* itself plus the mass attached to all the subsets of *B:*

$$Bel(A) = \sum_{B \subset A} m(B)$$

In other words, *Bel(A)* is all the evidence that speaks in favour of *A*.

As mentioned before, DST has a very important characteristic which sets it apart from probability theory: if $Bel(h_1) < 1$, then the remaining evidence $1\text{-}Bel(h_1)$ does not necessarily refute h_1 (in probability theory we would have $h_2 = 1\text{-}h_1$ if the FoD was $\{H=h_1,h_2\}$). Thus, some of the remaining evidence might *plausibly* be assigned to (sets of) hypotheses that are subsets of or include *A*. This is represented by the *plausibility function*:

$$Pl(A) = \sum_{A \cap B \neq 0} m(B)$$

In other words, *Pl(A)* represents the maximum possible belief in *A* that *could* be achieved if it was known that the remaining uncertain information (caused by errors in the input data, incomplete data, contradictory evidence etc.) does not refute *A*. DST provides additional useful outputs which will be demonstrated later in the archaeological case study.

A BPN is a value m_n in the range "0" to "1" for each hypothesis "1..n" in the FoD. The single restriction is that $m_{1..n}$ must sum to "1" for each source of evidence. A function that calculates a BPN for a set of hypotheses and meets these criteria is called a *BPN mass distribution function*.

In the case discussed here, it is necessary to calculate a BPN for each source of evidence by using only two types of information available in the GIS data base: (a) a categorized raster map and (b) a vector points map with the locations of already discovered sites. The basic idea is to check for *differences in proportions* between the sample (locations with sites) and the population (all locations). These differences are used to quantify the strength of belief in each of the FoD's three hypotheses. Generally, there are many possibilities for calculating BPN mass distribution functions and it may take considerable effort to find the one that works best in a given situation (see *e.g.* Ejstrud 2005).

It was felt that the available IDRISI modelling tools, as used in Ejstrud's (2003; 2005) study, fell short of catering for such complex relationships and a new implementation of DST in GRASS GIS was developed for use in the Borkeld case study. To be fair, arbitrarily complex BPN mass distribution functions can be designed in any GIS that supports raster map algebra, but a set of dedicated tools will make model design and updating much more efficient. The GRASS DST tools estimate $m(h_1)$ stepwise using a statistical approach. For this to work, we must supply two GIS maps:

1. a vector points map *S* with the locations of known sites.
2. a categorized raster map *M* that represents the evidence (*e.g.* "height").

The second requirement means that continuous variables, such as "height", have to be broken up into a finite number of discrete categories (*e.g.* "0-10 m, 11-20 m, 21-30 m, ..."). A choice must be made for a plausible categorization.

For each category *C* in the input evidence map *M*, we then compare the overall cell percentage of *C* in *M* (*Perc(M)*) with the percentage of category *C* cells that contain a site from *S* (*Perc(S)*). The assumption is that if *Perc(M) > Perc(S) at a significant level* then the evidence in *M* and *S* supports the hypothesis "no site" for *all* cells of category *C* in *M* and if *Perc(M) < Perc(S)* it supports "site". If the difference is found to be of very high significance, the belief mass assigned to one of these two hypothesis tends towards "1", if it is very low, it tends towards "0".

Just how significant the difference in percentages is depends on (a) the magnitude of the difference (b) the size of the sample and (c) the number of cells in the entire region. This is catered for by calculating a z-statistic for each case. The z-statistic gives the probability of observing a difference in proportions as extreme

as *Perc(M)-Perc(S)* under the assumption that there is *no* significant difference between *Perc(M)* and *Perc(S)*. In effect, this gives the probability of making a Type I error when declaring it significant (see "Handling uncertainty" for details on how this affects BPN values). If the total number *n* of sites in *S* is greater than 30, this will be derived from a standard normal distribution, otherwise from a t-distribution with *n* degrees of freedom (see Rogerson 2001 on the geographic application of this test).

Often, one will not be restricted as to what category ranges to choose (*e.g.* 1, 5 or 10 m intervals for evidence "height"). If there are too many categories in *M*, the result will be an overfitting of the BPN function. A model based on this will not be able to generalize well; its predictions are very likely to be wrong when checked against new data. On the other hand, if there are extremely few categories, the procedure might not be able to find significant evidence that supports either hypothesis. As a safeguard, a chi-square test is run over all categories. This calculates the probability of making an error when declaring the overall set of percentage differences significant.

Optionally, one or more raster maps may be supplied to represent the spatial distribution of "no site" bias, such as sparsely surveyed areas and one or more vector point attributes in the sites map can be chosen to represent "site" bias, such as unfounded trust in the location or identification of a site.