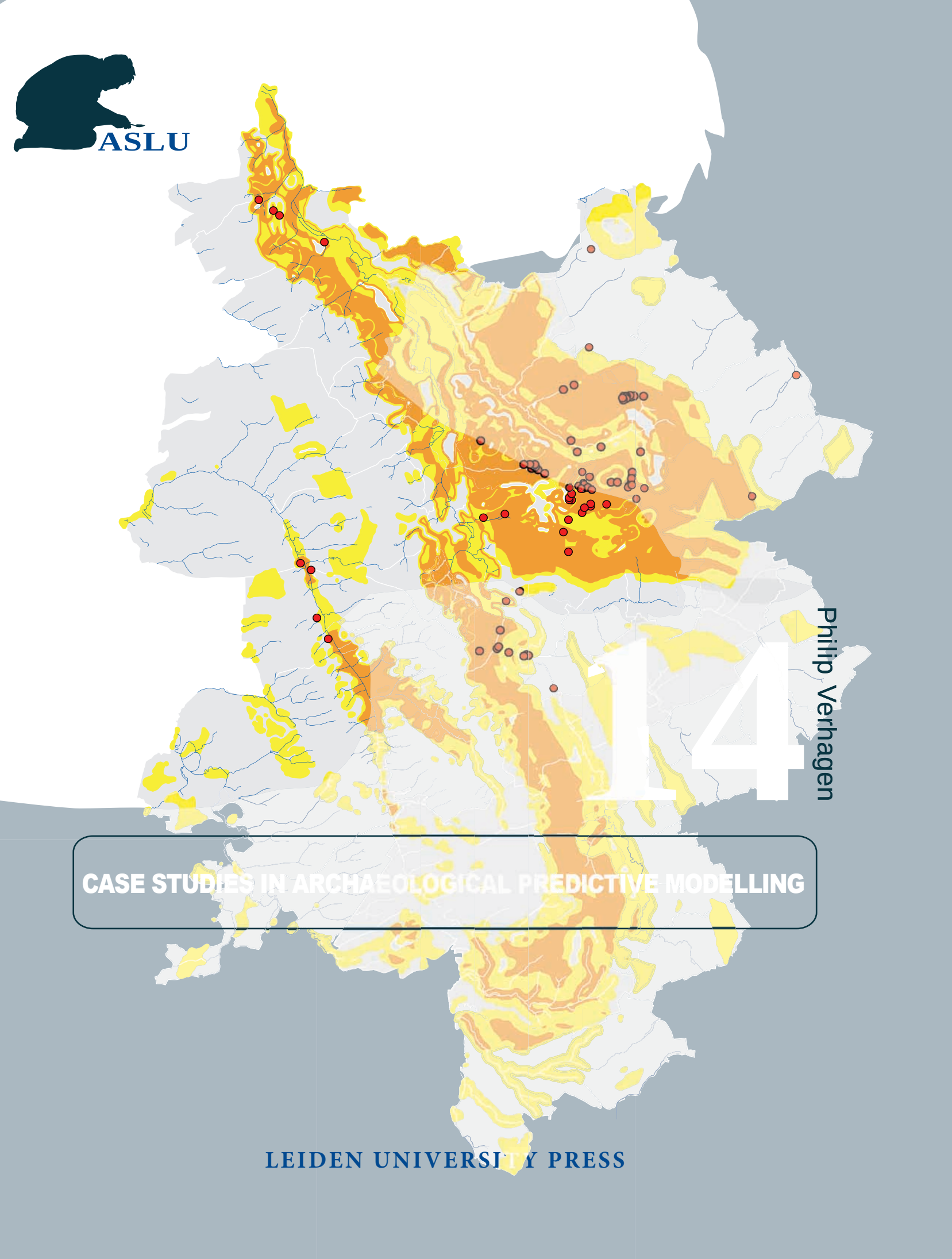
Case Studies in Archaeological Predictive Modelling

Verhagen, Jacobus Wilhelmus Hermanus Philippus; Bakels, C.C.; Kamermans, H.



ASLU

CASE STUDIES IN ARCHAEOLOGICAL PREDICTIVE MODELLING

Philip Verhagen

14

LEIDEN UNIVERSITY PRESS

CASE STUDIES IN ARCHAEOLOGICAL PREDICTIVE MODELLING

Leiden University Press

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

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

Cover illustration: Philip Verhagen
Cover design: Medy Oberendorff
Lay out: Philip Verhagen

ISBN 978 90 8728 007 9
NUR 682

ARCHAEOLOGICAL STUDIES LEIDEN UNIVERSITY

CASE STUDIES IN ARCHAEOLOGICAL PREDICTIVE MODELLING

PROEFSCHRIFT

ter verkrijging van
de graad van Doctor aan de Universiteit Leiden,
op gezag van de Rector Magnificus prof.mr. P.F. van der Heijden,
volgens besluit van het College voor Promoties
te verdedigen op woensdag 18 april 2007
klokke 16.15 uur

door

Jacobus Wilhelmus Hermanus Philippus Verhagen

geboren te Leiden
in 1966

Promotiecommissie

Promotores:
prof. dr. J.L. Bintliff
dr. H. Kamermans (co-promotor)

Referent:
prof. dr. G. Lock

Overige Leden:
prof. dr. H. Fokkens
prof. dr. W.J.H. Willems
prof. dr. J.C.A. Kolen
prof. dr. S.E. van der Leeuw
dr. P. van de Velde
dr. P.M. van Leusen

"Dit proefschrift is mede mogelijk gemaakt door RAAP Archeologisch Adviesbureau B.V."

TABLE OF CONTENTS

TABLE OF CONTENTS

TABLE OF CONTENTS

PREFACE

The core issue dealt with in this thesis is the improvement of the modelling techniques and testing methods used for creating archaeological predictive models. These models are made in the United States since the 1970s, and are used in Dutch archaeological heritage management since about 1990. The resulting maps predict the probability of the presence of archaeological remains in areas where no archaeological survey has been done. These predictions are based on an analysis of the location of known archaeological sites compared to factors like soil type or the proximity to water courses, and/or on hypotheses about the importance of these location factors. In the Netherlands, it is customary nowadays to use these maps in archaeological heritage management as a tool to decide whether archaeological survey is necessary or not. If the model predicts a low probability of the presence of archaeological remains, then survey will not be done. Apart from that, predictive maps can be used in environmental impact assessments. By creating a predictive model, a comparison can be made between proposed scenarios, e.g. for road building, and the option that is least damaging to the archaeological record can be established.

Even though archaeological predictive maps are commonly accepted tools for archaeological heritage management in the Netherlands, and are easy to use, they are also seriously debated in archaeological science. This is related to the presumed quality of the maps. In practice, it turns out that the statistical and conceptual models used for creating predictive maps are often based on incomplete data sets and flawed theories about the factors that determine why archaeological sites are found in a particular location. In this thesis, many of the issues relevant to setting up and testing predictive models are addressed.

This thesis is the result of various research projects that were carried out in the years 1995 through 2005. In this period I have been in the service of RAAP Archeologisch Adviesbureau BV (before 1998 Stichting RAAP) as a specialist in Geographical Information Systems (GIS). In those ten years, both RAAP and Dutch public archaeology have gone through rapid and profound change. RAAP originally started as a project for unemployed archaeologists in 1985, under the wings of the University of Amsterdam. In those days, archaeological excavations in the Netherlands were only permitted under the license of universities or the ROB[1]. Additional employment for archaeologists could only be found by doing non-destructive research. In a relatively short period, RAAP developed the foundations of Dutch archaeological prospection, by applying core sampling, field survey and to a lesser extent geophysical survey in order to perform preliminary archaeological research for land-management projects. By 1995 RAAP had already grown into a professional company specialized in non-destructive archaeological research. It was by then the only commercial archaeological company in the Netherlands, and had extended its activities into Germany and participated in European Union-funded scientific research projects. However, only a few years later, RAAP had ceased to work abroad and was competing on the national archaeological market created in anticipation of the implementation of the Valletta Convention in Dutch legislation. Today, the liberalization of Dutch archaeology has led to a large number of archaeological companies, competing on all areas of archaeological research (see Eickhoff, 2005). RAAP also stood at the basis of the development and application of archaeological predictive modelling in Dutch archaeology. Predictive modelling has become RAAP's most successful contribution to

[1] Rijksdienst voor het Oudheidkundig Bodemonderzoek, the Dutch National Archaeological Service

desk-based assessments, and RAAP still produces predictive maps up to this day. However, the methods and data used have experienced important changes over the past fifteen years.

In a commercial environment, one is seldom free to pursue a topic of research for a longer period of time. Therefore, the papers are not connected as if they are part of an ongoing academic research program, and they are of varying lengths. Most chapters are accompanied by a short commentary, in which the relevance of the chapters' conclusions will be discussed in the light of current insights. The exceptions are chapter 7, which is a fresh contribution, and therefore cannot be judged yet with hindsight, and chapter 5, which is commented together with chapter 6, as these two chapters are closely connected.

Four of the papers presented here are the direct or indirect result of my involvement in the *Archaeomedes* project (van der Leeuw, 1998; chapters 2, 3, 9 and 10). These papers deal with the application of GIS and archaeological predictive modelling in France and Spain, and are separated from the other papers that focus on the Netherlands. They are also the result of the collaborative efforts of the various research teams involved in *Archaeomedes*. The remaining papers are directly or indirectly connected to 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)[2]. This project started as the result of a series of discussions on predictive modelling in the Netherlands by the so-called *Badhuis*-group (Wansleeben and Kamermans, 1999; Verhagen *et al.*, 2000). More background on this project is given in chapter 1. Some of these papers were written in collaboration with the participants in this project.

The papers are not presented in chronological order, but have been rearranged to provide a more logical reading order. After an introductory chapter on the background and history of predictive modelling, three blocks of papers can be distinguished. The first block (chapters 2, 3 and 4) contains papers concerned with practical applications of methods and techniques to set up predictive models. Chapter 2 is a relatively short and practical paper on the creation of a predictive model in the Argonne region in north-eastern France. The predictive model presented in the paper is relatively straightforward, and focuses on the necessity to use the weak spots in the model as guidelines for prospection. Chapter 3 deals with a predictive model made for the Tricastin and Valdaine areas in south-eastern France. Both areas were studied intensively during the *Archaeomedes* project (van der Leeuw, 1998), when large numbers of previously unknown, buried sites were found. This new data set provided a unique opportunity to find out if the *communis opinio* of French archaeologists concerning the location of archaeological sites in the area could be tested against the results of a predictive model based on the new data. Chapter 4 explores a new way of dealing with 'soft' and 'hard' data sets in predictive modelling. The potential of Bayesian statistical methods has been acknowledged for a long time as a means to reconcile 'subjective' and 'objective' reasoning (Buck *et al.*, 1996). As an added bonus, it provides techniques for specifying the uncertainty of a model, as well as for calculating thresholds for sufficient data collection. However, its application in GIS has long been hampered by the absence of suitable software, and the general complexity of the calculations involved. The paper, using a case study of the municipality of Ede in the central Netherlands, tries to develop a relatively simple Bayesian model, using multicriteria decision-making techniques to quantify the 'expert judgment' side of the model, and shows how an expert-judgment model might be improved by introducing the archaeological data set itself into the model.

The second block, formed by chapters 5, 6 and 7, concentrates on sampling as a means to obtain the necessary data to develop and test archaeological predictive models. Around 2001 serious doubts began to arise on the utility of core sampling for finding lithic scatters. This question led to a thorough investigation of

[2] this project is part of the research program *'Protecting and Developing the Archaeological-Historical Landscape in the Netherlands'* (BBO; Bloemers and Wijnen, 2001; Bloemers, 2002), financed by the Netherlands Organisation for Scientific Research (NWO).

core sampling as a prospection technique, and of the characteristics of the Dutch archaeological record which could be detected using core samples. The results of this study were published in Dutch (Tol *et al.*, 2004). Chapter 5 shortly introduces the problems associated with core sampling as a prospection technique, and in chapter 6 an attempt is made to view the results of the study in the broader perspective of field survey and trial trenching, and the consequences that using different prospection techniques can have for the resulting archaeological data set. Chapter 7, the largest chapter of this thesis, takes sampling further as it tries to investigate how one could test predictive models in a quantitative manner, using either old or new archaeological data. It also gives some serious warnings concerning the current use of predictive models in the Netherlands: without quantitative quality norms, the models will remain uncontrollable.

The third and smallest block of papers looks at alternative ways of predictive modelling. Chapter 8 describes a study in south-eastern Spain, where an attempt was made to reconstruct the agricultural territories of known settlements. Even though the reconstructions were not intended for use as a predictive model, it still forms a good example of the way in which GIS can be used to develop so-called deductive models. The resulting models are of course hypothetical reconstructions, but they can easily be used for comparison with the archaeological data set, and in this way can serve to find out if the reconstructions bear any similarity to reality. This is very similar to the approaches that have later been developed by Whitley (2004; 2005) into a theory of causality-based predictive modelling. Chapter 9 is a short introduction on the potential of land evaluation in archaeology. Again, it is not dealing directly with predictive modelling, as it only gives some general ideas on how to use land evaluation in an archaeological context. However, land evaluation is one of the techniques that is easily applicable for the development of deductive, and at the same time quantitative, predictive models. Chapter 10, finally, develops some new ideas on how to use socio-cultural variables in predictive modelling. These three chapters are looking towards the future: is it possible to combine the world of quantitative methods and analysis with the theories and hypotheses that archaeologists have concerning site location, without reducing archaeological reality to too deterministic rules of behaviour?

REFERENCES

Bloemers, J.H.F. and M.-H. Wijnen (eds.), 2001. *Bodemarchief in Behoud en Ontwikkeling: de conceptuele grondslagen*. Van Gorcum, Assen.

Bloemers, J.H.F., 2002. 'Past- and Future-Oriented Archaeology: Protecting and Developing the Archaeological-Historical Landscape in the Netherlands', in: Fairclough, G. and S. Rippon (eds.), *Europe's Cultural Landscape: archaeologists and the management of change*. EAC, Brussels, p. 98-96.

Buck, C.E., W.G. Cavanagh and C.D. Litton, 1996. *Bayesian Approach to Archaeological Data*. John Wiley & Sons Ltd., Chichester.

Eickhoff, M., 2005. *Van het land naar de markt. 20 jaar RAAP en de vermaatschappelijking van de Nederlandse archeologie (1985-2005)*. RAAP Archeologisch Adviesbureau, Amsterdam.

Kamermans, H. and M. Wansleeben, 1999. 'Predictive modelling in Dutch archaeology, joining forces', in: Barceló, J.A., I. Briz and A. Vila (eds.), *New Techniques for Old Times – CAA98. Computer Applications and Quantitative Methods in Archaeology*. BAR International Series 757. Archaeopress, Oxford, pp. 225-230.

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

Leeuw, S.E. van der (ed.), 1998. *The Archaeomedes Project. Understanding the Natural and Anthropogenic Causes of Soil Degradation and Desertification in the Mediterranean Basin.* Office for Official Publications of the European Communities, Luxembourg.

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

Verhagen, P., M. Wansleeben and M. van Leusen, 2000. 'Predictive Modelling in the Netherlands. The prediction of archaeological values in Cultural Resource Management and academic research', in: Harl, O. and S. Strohschneider-Laue (eds.), *Workshop 4 Archäologie und Computer 1999*. Forschungsgesellschaft Wiener Stadtarchäologie, Vienna, pp. 66-82. CD-ROM.

Whitley, T., 2004. 'Causality and Cross-purposes in Archaeological Predictive Modeling', in: Fischer Ausserer, A., 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 International Series 1227. Archaeopress, Oxford, pp. 236-239 and CD-ROM (17 pages).

Whitley, T., 2005. 'A Brief Outline of Causality-Based Cognitive Archaeological Probabilistic Modelling', in: Leusen, M. van and H. Kamermans (eds.), *Predictive Modelling for Archaeological Heritage Management: A research agenda*. Nederlandse Archeologische Rapporten 29. Rijksdienst voor het Oudheidkundig Bodemonderzoek, Amersfoort, pp. 123-137.

ACKNOWLEDGEMENTS

No work of science can be completed without the help and encouragement of colleagues. My thanks go to a number of people who have expressed interest in my work, collaborated with me and have stimulated me to put these papers together in a thesis. In particular Hans Kamermans who, apart from teaching me more than I ever wished to know about the endless variations of Mr. Tambourine Man, was the person who convinced me that it could be done; Sander van der Leeuw who, as the coordinator of the *Archaeomedes* project, sparked my interest for archaeological scientific research back in 1992; Jean-François (Jeff) Berger, for sharing his ideas on geoarchaeology and his belief in GIS as an archaeological research tool; the other members of the Archaeomedes Rhône Valley team that I collaborated with closely: especially Michiel Gazenbeek, François Favory and Laure Nuninger; two very special people in Barcelona that I worked with on the *Archaeomedes* and *Río Aguas* project: Roberto Risch and Sylvia Gili; the other members of the BBO-project: Martijn van Leusen, Jos Deeben, Daan Hallewas and Paul Zoetbrood; John Bintliff, who didn't raise an eyebrow at the idea of becoming the supervisor of an external PhD-student in a subject only remotely connected to Mediterranean survey; and finally, my employers at RAAP who were kind enough to consider my work valuable and pay for it as well: the former director of RAAP, Roel Brandt, and its current director, Marten Verbruggen; and of course Adrie Tol, who had to put up with more statistics than he ever dreamed of for the cause of good archaeological prospection. And last but not least, to Jacoline, who has a keener eye than I have for structuring presentations and papers, and in that way has done her part to improve this thesis as well.

CHAPTER 1 A Condensed History of Predictive Modelling in Archaeology[1]

"It may be objected that human beings are not entirely rational. This is true, but neither are they fools nor do they choose to do more work than is necessary" (Chisholm, 1962)

1.1. INTRODUCTION

In this chapter, a review of the background and history of archaeological predictive modelling is given. It takes into account the international context and focuses on developments in the Netherlands over the past 15 years. The chapter covers a number of subjects that are also discussed in various other publications (a.o. Kohler, 1988; Dalla Bona, 1994; Verhagen *et al.*, 2000; van Leusen, 2002; Wheatley and Gillings, 2002; van Leusen and Kamermans, 2005; Verhagen *et al.*, 2005). In the context of this thesis however it is necessary and useful to restate the basic issues, and to bring the reader up to date with the latest developments.

Predictive modelling is a technique that, at a minimum, tries to predict "the location of archaeological sites or materials in a region, based either on a sample of that region or on fundamental notions concerning human behaviour" (Kohler and Parker, 1986:400). Predictive modelling departs from the assumption that the location of archaeological remains in the landscape is not random, but is related to certain characteristics of the natural environment. The precise nature of these relations depends very much on the landscape characteristics involved, and the use that prehistoric people may have had for these characteristics; in short, it is assumed that certain portions of the landscape were more attractive for human activity than others. If, for example, a society primarily relies on agricultural production, it is reasonable to assume that the actual choice of settlement location is, among others, determined by the availability of suitable land for agriculture.

The reasons for wanting to produce a predictive model for archaeology are very practical: when time and money do not allow a complete archaeological survey of an area, a predictive model can serve as a tool for the selection of the areas that are most likely to contain the archaeological phenomena of interest. Survey will then concentrate on these zones, and a maximum return on investment is obtained. This situation is commonly encountered in Cultural Resource Management (CRM), where archaeologists are forced to decide what to investigate within the constraints of tight budgets and time schedules, but it may also be an issue in an academic context, where the efficient expenditure of resources available during a fieldwork season can be an important aspect of scientific research. The designation of archaeologically important zones by means of predictive modelling can also be used to try to convince politicians and developers to choose the areas with the least 'archaeological risk' for their plans.

In most publications, two different approaches to the creation of predictive models can be distinguished. These have been referred to as 'inductive' and 'deductive' (Kamermans and Wansleeben, 1999), or as 'correlative' and 'explanatory' (Sebastian and Judge, 1988), but they are probably most adequately described as 'data driven' and 'theory driven' (Wheatley and Gillings, 2002). In the data driven approach,

[1] a slightly modified version of this chapter, together with the epilogue to this thesis, is published under the title 'Whither archaeological predictive modelling?' in W. Börner and S. Uhrlitz, 2006: *Workshop 10 Archäologie und Computer. Kulturelles Erbe und Neue Technologien. 7.-10. Novmeber 2005.* Stadtarchäologie Wien, Vienna (CD-ROM). The text was prepared by myself, but it is presented as a joint paper of the BBO Predictive Modelling project. Hans Kamermans and Martijn van Leusen are therefore both mentioned as co-authors in this published version.

statistical tests are applied to see if a relationship can be found between a sample of known archaeological sites and a selection of landscape characteristics (the 'environmental factors'). The correlations found are then extrapolated to a larger area. The theory driven approach starts by formulating a hypothesis on the location preferences of prehistoric people, and selecting and weighing the appropriate landscape parameters. The often cited ‘dichotomy’ between data driven and theory driven modelling, while useful for describing the different approaches to predictive modelling at a methodological level, ignores the fact that on the one hand the selection of data sets for inductive modelling is always theory laden, and that on the other hand the formulation of hypotheses of site location is always based on knowledge gathered from existing data. Elements of both approaches can therefore be found in many predictive modelling studies (Verhagen *et al.*, 2000; see also chapter 4).

Some authors have claimed that predictive modelling is a tool that can also be used to better understand the relationships between human activity and the natural environment, and as such may also serve a purely scientific purpose (Kamermans and Wansleeben, 1999). However, positioning predictive modelling as a truly scientific archaeological research tool somewhat misrepresents the issue. Site location analysis (by means of GIS and statistical methods) may result in new insights into site placement processes, and one can obviously extrapolate site location theories to see if they bear any similarity to the observed site patterns. However, this scientific approach implies that site location models are nothing more than tools to construct and verify hypotheses, whereas predictive modelling should result in a reliable estimate of the probability of encountering archaeological sites outside the zones where they have already been discovered in the past. So, while site location analysis and the construction of hypothetical site location models may be valuable contributions to the scientific process in themselves, they can only become *predictive* models if they are consciously designed as decision making tools.

1.2. THE ORIGINS OF ARCHAEOLOGICAL PREDICTIVE MODELLING

The roots of archaeological predictive modelling can be traced back to the late 1960s and the New Archaeology movement. The development of settlement pattern studies, initiated in American archaeology by Willey (1953; 1956; see Kohler, 1988), led many archaeologists to understand that settlement location is mainly determined by environmental factors. This ‘ecological’ approach was given theoretical backing by the introduction of geographical location theory in archaeology inspired by Chisholm (1962), who adapted the concepts laid out by Isard (1956). Chisholm's influential volume was followed some years later by the introduction of site catchment theory (Higgs and Vita-Finzi, 1972), which in essence tried to capture the rules that determine human spatial behaviour, approached from the angle of subsistence economy. The late 1960s also experienced a growing interest in the application of quantitative approaches for the analysis of site and settlement patterns, which gathered further momentum in the 1970s and eventually led to a large number of papers on sampling in archaeology (see e.g. Mueller, 1975), and the publication of two influential volumes on spatial analysis in archaeology (Hodder and Orton, 1976; Clarke, 1977).

Cultural Resource Management by that time had become an important issue in American archaeology, following the introduction of the National Historic Preservation Act in 1966. Federal agencies, confronted with the question how to deal with their responsibility to “identify historic properties on their lands (…) and to record such properties when they must be destroyed” (King, 1984), generated a demand for what was initially called ‘predictive survey’. The techniques developed by the Southwestern Archaeology Research Group (SARG) involved the comparison of expected to observed site distributions, and eventually laid the

foundations for data driven predictive modelling. Even though the term ‘predictive model’ can be traced back to some publications of the early 1970s, it is only in the second half of the 1970s that predictive models began to be produced on a larger scale in the United States. At first no specific methodology or product was favoured (Kohler, 1988).

By the late 1970s all the building blocks needed for data driven predictive modelling had been developed, and when computer technology became sufficiently advanced to allow for more sophisticated cartographic modelling by means of GIS, it was only a matter of following the leads provided. Initially, a lot of energy was devoted to the development of statistical and spatial analysis techniques for data driven predictive modelling, in which the work of Kenneth Kvamme has been most significant (Kvamme, 1983; 1984; 1988). GIS-based data driven modelling was already used in the United States as early as the mid 1980s, and the foundations of the 'American way' of predictive modelling are laid out in a number of publications, the most influential of which are Kohler and Parker (1986) and Judge and Sebastian (1988). By then, the methodology had fully developed, allowing Warren (1990) to write an easy to use 'recipe' on how to apply logistic regression to obtain the statistical correlations and predictions sought for. As is demonstrated by a number of applications found in Wescott and Brandon (2000) and Mehrer and Wescott (2006), this is still a commonly applied methodology in the United States. However, Altschul *et al.* (2004) note that the popularity of predictive modelling for land management purposes in the United States has declined over the past decade or so, because of the inability of the models to identify *all* archaeological resources: “the logic underlying this line of thought was that the agency would spend money up front to create an objective and verifiable model whose predictions would then substitute for large-scale survey” (Altschul *et al.*, 2004:5).

Theory driven modelling has always been a less practiced and accepted methodology for creating predictive models. The first published example of an explanatory predictive model using computer simulation is found in Chadwick (1978). Doorn (1993) arrived at a general explanation of settlement location in a study area in NW Greece with only three simple scenarios. His site location models took into account a limited number of variables that were manipulated differently for each scenario. In Doorn’s study, four variables were considered: communication, safety, availability of water and quality of agricultural land. For each scenario, these factors were rated differently: a self-sufficient economy will place greater emphasis on the presence of sufficient water and land, whereas a community that employs a defensive strategy will place safety first. The attraction of these theory driven predictive models (which were later more thoroughly explored by Whitley (2004; 2005)) lies in their ease of use and in the ability to contrast them with known settlement patterns in order to generate hypotheses concerning the actual location preferences. At the same time, these advantages are also the dangers: the models can be highly speculative and may give rise to spurious explanations of site location from a limited environmental or economic perspective. They also still need to be compared to an archaeological data set in order to be tested. If this data set is biased, then the validity of a theory driven model cannot be established.

1.3. GIS IN ARCHAEOLOGY

The early 1990s were characterized by a 'boom' in archaeological GIS applications. GIS was a huge success, and has resulted in the publication of various volumes of archaeological applications (Allen and Zubrow, 1990; Lock and Stančič, 1995; Lock, 2000; Wheatley and Gillings, 2002); it has filled large portions

of the proceedings of CAA[2]-conferences up to this day. Predictive modelling has always been a very important issue in archaeological GIS use. In fact, many archaeologists in the United States even seem to conflate the two (Altschul *et al.*, 2004). Looking back, the success of GIS can largely be attributed to its ability to communicate both its concepts as well as its results to the archaeological community. Archaeology is a social science, and 'hard science' techniques and methods have always been regarded with a bit of suspicion; statistical methods and theory are not well understood by a majority of archaeologists. Compared to the 'difficult' statistical methods, GIS is a relatively simple technique that can be used in a meaningful way without having to understand the mathematical basis of it[3]. Of course, this also carries with it the danger of over-simplification of complex geographical problems, which has resulted in quite a number of not so sophisticated archaeological GIS applications, that were more inspired by the software's proverbial map overlay capabilities than by sound archaeological research questions.

Another aspect of the success of GIS is the fact that cartographic output produced by GIS software is easy to understand and can convey a convincing image of the results of the analyses performed. At the downside, these results can be made to 'look good' by means of the software's inbuilt cartographic tools, thereby obscuring the fact that the map shown is the result of any number of manipulations of the underlying data sets[4].

However, by the time GIS developed into an important tool for both archaeological research and data management, New Archaeology had fallen out of grace with the scientific community. This is not the place to go into detail about the rise of post-processual archaeology, but it can be said that much of the academic debate in the 1990s concerning predictive modelling is based on the dichotomy between the processual and post-processual way of reasoning. Basically, the accusation of reductionism has been the main thread of academic criticism on predictive modelling - and of many other attempts in archaeology to introduce a more quantitative, 'hard-science' approach to archaeological questions. If we look at the published criticism of GIS in archaeology in the mid 1990s, the main concern of post-processual archaeologists was that the use of GIS (and therefore predictive modelling) constituted a regression to the days of New Archaeology, and re-introduced the now abandoned ideas of environmental determinism and site catchment theory (Gaffney and van Leusen, 1995; Wheatley, 1996; Wansleeben and Verhart, 1997).

Part of the environmental focus of predictive modelling and GIS in archaeology is a direct consequence of the way in which GIS originated outside archaeology, and of the environmental data sets that have become available in digital form. GIS was primarily designed as a software tool to analyse land use, and environmental questions were therefore among the first to be tackled[5]. It is not surprising that the social sciences were somewhat later in adopting it as a tool for geographical analysis and representation, and had to find their way in it. Apart from that, the demise of New Archaeology had not yet led to a new theory and methodology for archaeological spatial analysis. The well-known book 'A phenomenology of landscape' by Tilley (1994) for example, which lays the foundations of a post-processual theory of space, is conspicuously devoid of maps. The late 1990s were therefore characterized by several attempts to include the less tangible aspects of spatial behaviour into the archaeological application of GIS (most notably the development of viewshed analysis; Gaffney *et al.*, 1995; Wheatley, 1995; Llobera, 1996; Wheatley and Gillings, 2000; van Leusen, 2002). This development has however been confined to the academic community, and up to this day

[2] the annual conference of Computer Applications and Quantitative Methods in Archaeology
[3] although, in the early days, the manipulation of GIS-software required quite some expertise in computer science
[4] however, when looking at many archaeological GIS-presentations at conferences and the resulting published papers, it is amazing to see how little advantage is taken of this very powerful capacity of the software.
[5] ESRI, the producer of ARC/INFO and one of the world's largest GIS companies, is an acronym for Environmental Systems Research Institute

has had very little influence on predictive modelling practice in public archaeology. Furthermore, it seems to have been completely ignored in the United States.

1.4. THE CONTROVERSY ON PREDICTIVE MODELLING

While the heat of the GIS-debate has gradually subsided, predictive modelling is still a controversial issue, and authors like David Wheatley have consistently and actively criticized its application up to this day. The main pitfalls of (data driven) predictive modelling, signalled in various publications (e.g. Wansleeben and Verhart, 1997; Wheatley and Gillings, 2002; van Leusen *et al.*, 2005), are:

- the use of incomplete archaeological data sets;
- the biased selection of environmental parameters, often governed by the availability of cheap data sets such as digital elevation models;
- as a consequence, a neglect for the influence of cultural factors, both in the choice of environmental parameters, as well as in the archaeological data set;
- and lastly, a neglect of the changing nature of the landscape

Note that the problems mentioned are all related to the inability of archaeologists to obtain the appropriate data sets needed for predictions that cover all aspects of site location. While it is certainly true that many published predictive models are simplistic at an explanatory level, the real issue is whether full explanatory power is actually a necessary characteristic of a good predictive model. If the model works at the practical level and correctly assigns archaeological sites to zones of high probability, then explanation could perhaps be of secondary importance. However, as most data driven models use a selection of data that is biased to the natural environment, they implicitly represent an explanatory, environmental deterministic model that is far from covering all factors that determine site location (see e.g. Gaffney and van Leusen, 1995; Ebert, 2000). Even staunch advocates of data driven modelling admit that their models work better with prehistoric communities that highly relied on the natural environment for their subsistence, like hunter-gatherers, than for societies that developed more complex cultural systems. However, data driven models can perfectly well be made by adding cultural parameters to the usual set of environmental factors (see also chapter 10; Ridges, 2006), and theory driven models can be based on flawed theories of site location. The way forward therefore is to look for a combination of both approaches.

1.5. PREDICTIVE MODELLING IN CULTURAL RESOURCE MANAGEMENT

By the end of the 1990s, GIS had more or less split the archaeological community into two camps: the academic acceptance of GIS had been relatively slow and was accompanied by serious doubts about its usefulness as a tool for scientific analysis (Wheatley, 2003). In public archaeology on the other hand, GIS had been embraced as a convenient tool to combine geographical data with database management systems in order to store and retrieve the enormous amounts of available archival information. Many national and regional archaeological authorities made the step somewhere in the 1990s to enter their paper archives into a GIS-based database management system (e.g. Roorda and Wiemer, 1992; Guillot and Leroy, 1995; Blasco *et al.*, 1996; see García Sanjuán and Wheatley, 1999, for a comprehensive overview), that can be used for a quick retrieval of information, for example to judge whether planned developments should be accompanied by archaeological

research. In the United States, Canada and the Netherlands, predictive modelling has been an integral part of this development. The predictive map was seen as a powerful instrument to draw the attention of local governments and developers to the archaeological potential of an area, and to quantify the risks that could be run when development plans were left unchecked for archaeological 'problems'. In other European countries there was no such development until very recently (Ducke and Münch, 2005; Ejstrud, 2003), with the exception of Slovenia, where predictive maps were developed in the late 1990s (Stančič and Kvamme, 1999; Stan?i? and Veljanovski, 2000; Stančič *et al.*, 2001). One of the most important objections against the use of predictive models in CRM is given by Wheatley (2003): the self-fulfilling nature of the predictions made, as these are used to decide where to do intensive prospection in the case of development plans. In the United Kingdom and France, full scale prospection is therefore customarily performed when development plans are in the initial stages and the archaeological risks need to be established. However, this means that there is very little opportunity to influence planning decisions *before* they come from the drawing board, other than by using the existing sites and monuments records, whereas in the Netherlands some influence of predictive mapping on the (political) decisions made can be observed (e.g. in the case of environmental impact assessments; Scholte Lubberink *et al.*, 1994). However, even in the Netherlands it is surprising to observe that very little effort has gone into the quantification of the risks involved, both in terms of the archaeological value of the areas threatened as well as in terms of the amount of money that 'cleaning up' of the archaeological problem might cost.

1.6. PREDICTIVE MODELLING IN THE NETHERLANDS

In 1997, the Netherlands witnessed the birth of the first predictive map on a national scale, the *Indicatieve Kaart van Archeologische Waarden*[6] (IKAW; Deeben *et al.*, 1997). Looking back, this has been a decisive moment for predictive modelling in the country. Until then, predictive maps were only made by RAAP on a local scale, and reflected the American way of predictive modelling (see Kamermans and Wansleeben, 1999). It was noted earlier by Brandt *et al.* (1992) and later by van Leusen (1996) that this method was not well suited for application in the Dutch archaeological context, mainly because of the lack of reliable archaeological data. Brandt *et al.* (1992) attributed this to the peculiar nature of Dutch archaeology, of which much is hidden underneath the soil, thereby effectively ruling out field walking as a cheap solution for checking the model. However, it also has to be said that no attention has been paid to alternative solutions, like analyzing the prospections that were done and correcting them for possible biases.

The producers of the IKAW ignored most of the methodological problems signalled earlier and produced a quantitative map on the basis of demonstrably biased archaeological data and a limited set of environmental variables (soil type and groundwater table, later extended with geological information in the Holocene part of the Netherlands). It was therefore no surprise that the map was greeted with quite a bit of scepticism. Several regions could easily be identified as having the 'wrong prediction', mainly because of a lack of archaeological data. This was partly circumvented by consulting experts on specific regions and archaeological periods, which led to several adaptations to the original map (Deeben *et al.*, 1997). In 2002, a second version was released (Deeben *et al.*, 2002). However, the IKAW is not a very accurate map, and it was therefore advised to use it only in the initial stages of development plans, preferably at the provincial or national level, rather than as a guide of where to do survey or to prepare mitigating measures.

---

[6] *Indicative Map of Archaeological Values*

RAAP's answer was to stop using data driven modelling as a method to produce predictive maps. Instead, predictions are now made by them and other archaeological companies using a more deductive and intuitive strain of reasoning, by asking which features in the landscape would have been attractive for settlement in a specific archaeological period. In order to perform this type of modelling, a thorough knowledge is required of the geo(morph)ology of an area, and it turned out that using this knowledge can be very valuable for predictive mapping, certainly when it is combined with geo-archaeological prospection by means of core sampling. The resulting maps can best be seen as archaeological interpretations of the landscape, giving each geomorphological element an archaeological value. No quantitative analysis at all is involved, although in some cases the archaeological data set is used as an independent check, to see whether it confirms the map, and if not, why there is a discrepancy.

It was around the time that the IKAW was first published that contacts were established between academic researchers (Hans Kamermans, Martijn van Leusen, Milco Wansleeben and Harry Fokkens) who had shown an interest in predictive modelling, but never actually had participated in the production of these maps for use in public archaeology, and the producers of predictive maps at RAAP (Philip Verhagen and Eelco Rensink) and the ROB (Ronald Wiemer, Jan Kolen and Jos Deeben). In a series of discussions it became obvious that even though the producers of predictive maps perfectly understood the problems identified by the academic community, they were unable to tackle these by themselves, basically because of a lack of research money and communication on both sides. Where for the 'academics' it sufficed to signal a flaw in the modelling procedures, perhaps try out a new technique, and then move on to a new subject, the people working in public archaeology were unable to pick up these insights and convert them into practical working solutions. Furthermore, the 1997 IKAW-paper was about the first one that was published in which a broader audience could judge the methods and choices made for this particular model. The results of these so-called *Badhuis*-discussions were published by Kamermans and Wansleeben (1999) and Verhagen *et al*. (2000). These papers set the agenda for the grant application that led to the establishment of 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). In this project, academic (Hans Kamermans and Martijn van Leusen) and non-academic (Jos Deeben, Daan Hallewas, Paul Zoetbrood and Philip Verhagen) researchers joined forces with the specific aim in mind to explore the possibilities for methodological improvement and greater efficiency of predictive models in Dutch and international practice.

The reason that a predictive modelling research project could be funded in 2002, whereas money for this subject had been difficult to find before (with the exception of the very first models made by RAAP in 1990, which were partly funded by NWO[7]), was the implementation of the Valletta Convention that was rapidly changing the face of Dutch public archaeology. From 1998 on, in anticipation of the revision of the Law on Ancient Monuments of 1988 (now scheduled for early 2006), the Dutch government decided to change the structure of archaeological heritage management, by gradually allowing commercial excavation and creating an archaeological market, which by now has led to the establishment of some 50 archaeological companies in the Netherlands. In order to have a well-functioning process of archaeological heritage management, a system of quality norms was designed (*Kwaliteitsnorm Nederlandse Archeologie* or KNA; College voor de Archeologische Kwaliteit, 2001), and predictive modelling was by then so well embedded in the archaeological working process, that it became an obligatory step in archaeological desk top studies to consult predictive maps, or to create them if necessary. At the same time, money was invested in the

[7] the Netherlands Organization for Scientific Research, which is responsible for the allocation of government funds for scientific research

archaeological research programme '*Bodemarchief in Behoud en Ontwikkeling*' [8] or BBO (Bloemers, 2002) that took the protection of the archaeological heritage to heart, and aimed at developing better methods to ensure its protection. Predictive modelling, as one of the methods to achieve this goal, therefore was recognized as an integral and indispensable part of the BBO programme.

1.7. THE BBO PREDICTIVE MODELLING PROJECT

The BBO predictive modelling project started out with the preparation of a baseline report on the current state of affairs in predictive modelling, both in the Netherlands and internationally (van Leusen *et al.*, 2005). The baseline report contained a comprehensive overview of all the issues relevant to predictive modelling, many of which have been mentioned in the preceding sections of this chapter. Following this state of the art, the report focused on the six major themes that are most likely to yield significant improvements on current predictive modelling practice in the Netherlands. These are:

- *The quality of the archaeological input data.* While it is generally recognized that the existing archaeological site databases are not representative of the total archaeological record, little effort has gone into the improvement of the currently available data for the purpose of developing better predictive models. Suggested improvements include the development of specific data collection programs, and the analysis of existing archaeological data in order to identify the discovery and research processes that have produced the 'official record'.
- *Environmental input factors.* More detailed mapping, e.g. by LIDAR-based elevation models or high-resolution palaeo-geographic research will result in more precise zonations. A better understanding of post-depositional processes is necessary not only to predict the location, but also the quality of the archaeological remains.
- *The inclusion of socio-cultural factors.* Socio-cultural factors are virtually absent in predictive modelling studies up to now, and methods for using them in a predictive modelling context will have to be developed more or less from scratch.
- *Higher spatial and temporal resolution.* Predictive models for different archaeological periods are needed, as well as more detailed maps than currently are available.
- *(Spatial) statistics*. The statistical toolbox currently used in predictive modelling is rather small and limited. Recent developments in statistics, like Bayesian inference, fuzzy logic, resampling or the application of geo-statistics have more or less passed by predictive modelling.
- *Testing.* The only way to perform quality control of predictive models is by means of testing. However, in practice tests that are carried out are limited, and no consensus exists as to the best way of testing.

In order to obtain a critical review of the baseline report, a two-day workshop was organized in Amersfoort in May 2003. Various experts in predictive modelling from the Netherlands and abroad were asked to give their view on the issues mentioned in the baseline report, and to present a position paper drawing on their own expertise. This meeting resulted in an edited volume of proceedings (van Leusen and Kamermans, 2005). The most important conclusions that can be drawn from the project results are:

[8] '*Protecting and Developing the Archaeological-Historical Landscape in the Netherlands*'

- the Netherlands occupy a unique position in Europe because of the way in which predictive models are used in archaeological heritage management;
- at the same time, the shortcomings of predictive modelling in the Netherlands, while generally recognized, are in practice only approached from the angle of improvement of the environmental data sets; the quantitative approach has clearly lost its appeal to the Dutch archaeological community, and the low quality of the IKAW is at least partly responsible for this development;
- at the same time, the development of a predictive model in Brandenburg (Germany) by Ducke and Münch (2005), shows that many of the shortcomings of especially the IKAW can be dealt with when it comes to generating and using higher quality archaeological and environmental input data sets;
- other authors have experimented with alternatives to the traditional, correlative methods prevalent in especially American predictive modelling; Whitley (2004; 2005) for example argues that a formalized theory driven modelling approach is not only more effective and scientifically valid, but also much cheaper than investing enormous amounts of time and money in the collection and analysis of the data needed for data driven modelling; in his view, archaeological data analysis, using relatively modest sample sizes, can come after the modelling is done, and will only serve as a confirmation or refutation of the modelling results[9];
- predictive modellers who nevertheless want to stick to the data driven line of modelling can now use more sophisticated statistical methods than even a few years ago; especially Bayesian inference (see chapter 4; Millard, 2005) and Dempster-Shafer theory (Ejstrud, 2003; 2005) are serious competitors to the currently available traditional statistical methods, as both are able to formalize expert judgement into a quantitative framework; both these methods were tested out in January 2005 in a workshop in Amsterdam, and the results of the test will be published in the final volume of the BBO-project.
- testing of predictive models is badly needed, as it is the only objective means to assess the quality of the models (see chapter 7); testing implies data collection and analysis within a probabilistic sampling framework in which low probability zones are also surveyed; this implies a break with current practice, which mainly limits survey to high probability zones.

While the BBO-project is certainly the most conspicuous effort made for the improvement of predictive modelling over the past fifteen years, it has to be pointed out that a cautious reassessment of predictive modelling also seems to be going in the United States. A GIS conference in March 2001 in Argonne, Illinois (Mehrer and Wescott, 2006), aimed at discussing many of the issues mentioned in this chapter. However, there is little evidence that this has already resulted in changes in North-American practice, with the exception of the work done by Whitley (2004; 2005; see also Altschul *et al.*, 2004). In Europe, predictive modelling slowly seems to gain credit as a useful method for archaeological heritage management. Several regional predictive maps have been developed in the past five years or so, in countries like Germany, France, Denmark and the Czech Republic, and other countries are considering to prepare predictive maps. However, most of these efforts are not very accessible to the outside world, with the clear exception of the *Archäoprognose Brandenburg* in Germany (Ducke and Münch, 2005).

[9] and by setting up a range of models, based on different explanatory frameworks that may include many site location factors and different weighting of these factors, the best performing model can be selected with relative ease

So, predictive modelling is now standing at a crossroad: will we continue to make predictive models like we have been doing for more than 15 years now, or is it time to adopt a different approach? In the epilogue to this thesis, I will try to answer this question.

REFERENCES

Allen, K.M.S., S.W. Green, and E.B.W. Zubrow (eds.), 1990. *Interpreting Space: GIS and Archaeology.* Taylor and Francis, New York.

Altschul, J.F., L. Sebastian and K. Heidelberg, 2004. *Predictive Modeling in the Military. Similar Goals, Divergent Paths.* Preservation Research Series 1. SRI Foundation, Rio Rancho (NM). http://www.srifoundation.org/pdf/FINALLEG.pdf, accessed on 24-11-2005.

Blasco, C., J. Espiago and J. Baena, 1996. 'The role of GIS in the management of archaeological data: an example of an application to the Spanish administration', in: Aldenderfer, M. and H. Maschner (eds.), *Anthropology, space, and geographic information systems.* Oxford University Press, New York, pp. 189-201.

Bloemers, J.H.F., 2002. 'Past- and Future-Oriented Archaeology: Protecting and Developing the Archaeological-Historical Landscape in the Netherlands', in: Fairclough, G. and S. Rippon (eds.), *Europe's Cultural Landscape: archaeologists and the management of change*. EAC, Brussels, pp. 89-96.

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.

Chadwick, A.J., 1978. 'A computer simulation of Mycenaean settlement', in: Hodder, I. (ed.), *Simulation studies in archaeology*. Cambridge University Press, Cambridge, pp. 47-57.

Chisholm, M., 1962. *Rural settlement and land use: an essay in location.* Hutchinson University Library, London.

Clarke, D.L., 1977. *Spatial archaeology.* Academic Press, London.

College voor de Archeologische Kwaliteit, 2001. *Kwaliteitsnorm Nederlandse Archeologie, versie 2.0*. Ministerie van Onderwijs, Cultuur en Wetenschappen, The Hague.

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

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: Willems, W., H. Kars and D. Hallewas (eds.), *Archaeological Heritage Management in the Netherlands. Fifty Years State Service for Archaeological Investigations*. Rijksdienst voor het Oudheidkundig Bodemonderzoek, Amersfoort, pp. 76-118.

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

Doorn, P.K., 1993. 'Geographical Location and Interaction Models and the Reconstruction of Historical Settlement and Communication: The Example of Aetolia, Central Greece.' *Historical Social Research*, 18:22-35.

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

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

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

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

Gaffney, V. and M. van Leusen, 1995. 'Postscript - GIS, environmental determinism and archaeology: a parallel text', in: Lock, G. and Z. Stančič (eds.), *Archaeology and Geographical Information Systems: A European Perspective*. Taylor and Francis, London, pp. 367-382.

Gaffney, V., Stančič, Z. and H. Watson, 1995. 'The impact of GIS on archaeology: a personal perspective', in: Lock, G. and Z. Stančič (eds.). *Archaeology and Geographical Information Systems: A European Perspective*. Taylor and Francis, London, pp. 211-229.

García Sanjuán, L. and D. Wheatley, 1999. 'The state of the Arc: differential rates of adoption of GIS for European Heritage Management'. *European Journal of Archaeology* 2:201-228.

Guillot, D. and G. Leroy, 1995. 'The use of GIS for archaeological resource management in France: the SCALA project, with a case study in Picardie', in: G. Lock and Z. Stančič (eds.), *Archaeology and Geographical Information Systems: A European Perspective*. Taylor and Francis, London, pp. 15-26.

Higgs, E.S. and C. Vita-Finzi, 1972. 'Prehistoric economies. A territorial approach', in: Higgs, E.S. (ed.), *Papers in economic prehistory*. Cambridge University Press, Cambridge, pp. 27-36.

Hodder, I. and C. Orton, 1976. *Spatial analysis in archaeology*. New Studies in Archaeology, Vol. 1. Cambridge University Press, Cambridge.

Isard, W., 1956. *Location and space economy: a general theory relating to industrial location, market, land use, trade and urban structure.* The Regional Science Studies Series, Vol. 1. The M.I.T. Press, Cambridge (Massachusetts).

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

Kamermans, H. and M. Wansleeben, 1999. 'Predictive modelling in Dutch archaeology, joining forces', in: Barceló, J.A., I. Briz and A. Vila (eds.), *New Techniques for Old Times – CAA98. Computer Applications and Quantitative Methods in Archaeology*. BAR International Series 757. Archaeopress, Oxford, pp. 225-230.

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

King, T.F., 1984. 'An Overview of the Archaeological Law at the Federal Level'. *American Archaeology* 4:115-118.

Kohler, T.A., 1988. 'Predictive Modelling: History and Practice', in: Judge, J.W. and L. Sebastian (eds.), *Quantifying the Present and Predicting the Past: Theory, Method and Application of Archaeological Predictive Modelling.* U.S. Department of the Interior, Bureau of Land Management, Denver, pp. 19-59.

Kohler, T.A. and S.C. Parker, 1986. 'Predictive models for archaeological resource location', in: Schiffer, M.B. (ed.), *Advances in Archaeological Method and Theory, Vol. 9*. Academic Press, New York, pp. 397-452.

Kvamme, K.L., 1983. *A manual for predictive site location models: examples from the Grand Junction District, Colorado.* Bureau of Land Management, Grand Junction District.

Kvamme, K.L., 1984. 'Models of Prehistoric Site Location near Pinyon Canyon, Colorado', in: Condie, C.J. (ed.), *Papers of the Philmont Conference on the Archaeology of Northeastern New Mexico.* Proceedings of the New Mexico Archaeological Council 6(1), Albuquerque.

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

Leusen, P.M. van, 1996. 'Locational Modelling in Dutch Archaeology', in: Maschner, H.D.G. (ed.), *New Methods, Old Problems: Geographic Information Systems in Modern Archaeological Research.* Occasional Paper no. 23. Center for Archaeological Investigations, Southern Illinois University, Carbondale, pp. 177-197.

Leusen, P.M. van, 2002. *Pattern to Process. Methodological investigations into the formation and interpretation of spatial patterns in archaeological landscapes.* Rijksuniversiteit Groningen, Groningen. PhD thesis.

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

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

Lock, G. (ed.), 2000. *Beyond the Map. Archaeology and Spatial Technologies.* NATO Science Series, Series A: Life Sciences, vol. 321. IOS Press / Ohmsha, Amsterdam.

Lock, G. and Z. Stančič (eds.), 1995. *GIS and archaeology: a European perspective.* Taylor and Francis, London.

Llobera, M., 1996. ‘Exploring the topography of mind: GIS, social space and archaeology’. *Antiquity* 60:612-622.

Mehrer, M. and K. Wescott (eds.), 2006. *GIS and Archaeological Site Location Modeling.* CRC Press, Boca Raton.

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

Mueller, J.W. (ed.), 1975. *Sampling in archaeology.* University of Arizona Press, Tucson.

Ridges, M., 2006. ‘Understanding H-G behavioural variability using models of material culture: An example from Australia’, in: Mehrer, M., and K. Wescott (eds.), *GIS and Archaeological Site Location Modeling*. CRC Press, Boca Raton, pp. 123-143.

Roorda, I.M. and R. Wiemer, 1992. ‘Towards a new archaeological information system in the Netherlands’, in: Lock, G. and J. Moffett (eds.), *Computer Applications and Quantitative Methods in Archaeology 1991*. BAR International Series (S577). Tempus Reparatum, Oxford, pp. 85-88.

Sebastian, L. and W.J. Judge, 1988. ‘Predicting the Past: Correlation, Explanation and the Use of Archaeological Models’, in: Judge, J.W. and L. Sebastian (eds.), *Quantifying the Present and Predicting the Past: Theory, Method and Application of Archaeological Predictive Modelling.* U.S. Department of the Interior, Bureau of Land Management, Denver, pp. 1-18.

Scholte Lubberink, H.B.G., J.W.H.P. Verhagen and H. van Londen, 1994. *Archeologisch onderzoek ten behoeve van de trajectstudie / m.e.r. Rijksweg 4: Kruithuisweg (Delft) – Kethelplein (Schiedam).* RAAP-rapport 94. Stichting RAAP, Amsterdam.

Stančič, Z. and K. Kvamme, 1999. ‘Settlement patterns modelling through Boolean overlays of social and environmental variables’, in: Barceló, J.A., I. Briz and A. Vila (eds.), *New Techniques for Old Times – CAA98. Computer Applications and Quantitative Methods in Archaeology*. BAR International Series 757. Archaeopress, Oxford, pp. 231-238.

Stančič, Z. and T. Veljanovski, 2000. ‘Understanding Roman settlement patterns through multivariate statistics and predictive modelling’, in: Lock, G. (ed.), *Beyond the Map. Archaeology and Spatial Technologies.* NATO Science Series, Series A: Life Sciences, vol. 321. IOS Press / Ohmsha, Amsterdam, pp. 147-156.

Stančič, Z., T. Veljanovski, K. Oštir and T. Podobnikar, 2001. ‘Archaeological Predictive Modelling for Highway Construction Planning’, in: Stančič, Z. and T. Veljanovski (eds.), *Computing Archaeology for Understanding the Past - CAA2000 - Computer Applications and Quantitative Methods in Archaeology, Proceedings of the 28th Conference*. BAR International Series 931, Archaeopress, Oxford, pp. 233-238.

Tilley, C., 1994. *A phenomenology of landscape. Places, paths and monuments.* Berg, Oxford.

Verhagen, P., M. Wansleeben and M. van Leusen, 2000. ‘Predictive modelling in the Netherlands. The prediction of archaeological values in Cultural Resource Management and academic research’, in: Hartl, O. and S. Strohschneider-Laue (eds.), *Workshop 4 Archäologie und Computer 1999*. Forschungsgesellschaft Wiener Stadtarchäologie, Vienna, pp. 66-82. CD-ROM.

Verhagen, P., J. Deeben, D. Hallewas, P. Zoetbrood, H. Kamermans, and M. van Leusen, 2005. 'A review of Predictive Modelling for Archaeological Heritage Management in the Netherlands', in: Berger, J.-F., 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*. Éditions APDCA, Antibes, pp. 83-92.

Wansleeben, M. and L.B.M. Verhart, 1997. ‘Geographical Information Systems. Methodical progress and theoretical decline?’ *Archaeological Dialogues* 4:53-70.

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

Wescott, K.L. and R.J. Brandon, 2000. *Practical Applications of GIS for Archaeologists. A Predictive Modeling Toolkit.* Taylor and Francis, London.

Willey, G.R., 1953. *Prehistoric settlement in the Virú Valley, Peru*. Bureau of American Ethnology Bulletin 155. United States Government Printing Office, Washington.

Willey, G.R. (ed.), 1956. *Prehistoric settlement pattern in the New World.* Viking Fund Publications in Anthropology 23. New York.

Wheatley, D. 1996. ‘Between the lines: the role of GIS-based predictive modelling in the interpretation of extensive survey data’, in: Kamermans, H. and K. Fennema (eds.), *Interfacing the Past. Computer applications and quantitative methods in Archaeology CAA95*. Analecta Praehistorica Leidensia 28:275-292.

Wheatley, D., 1995. ‘Cumulative viewshed analysis: a GIS-based method for investigating intervisibility, and its archaeological application’, in: Lock, G. and Z. Stančič (eds.), *GIS and archaeology: a European perspective*. Taylor and Francis, London, pp. 171-185.

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

Wheatley, D. and M. Gillings, 2000. 'Vision, perception and GIS: developing enriched approaches to the study of archaeological visibility', in: Lock, G. (ed.), *Beyond the Map. Archaeology and Spatial Technologies.* NATO Science Series, Series A: Life Sciences, vol. 321. IOS Press / Ohmsha, Amsterdam, pp. 1-27.

Wheatley, D. and M. Gillings, 2002. *Spatial technology and archaeology: the archeaological applications of GIS.* Taylor and Francis, London.

Whitley, T., 2004. 'Causality and Cross-purposes in Archaeological Predictive Modeling', in: Fischer Ausserer, A., 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 International Series 1227. Archaeopress, Oxford, pp. 236-239 and CD-ROM (17 pages).

Whitley, T., 2005. 'A Brief Outline of Causality-Based Cognitive Archaeological Probabilistic Modelling', in Leusen, M. van and H. Kamermans (eds.), *Predictive Modelling for Archaeological Heritage Management: A research agenda*. Nederlandse Archeologische Rapporten 29. Rijksdienst voor het Oudheidkundig Bodemonderzoek, Amersfoort, pp. 123-137.

PART 1 PRACTICAL APPLICATIONS

In this part, I have grouped three papers, each presenting a case study with different approaches to creating archaeological predictive models.

In chapter 2, I will discuss a predictive model made for the Argonne region in north-eastern France. The aim of this model was to predict the location of pottery kiln sites that were used for the mass production of ceramics in the Roman period. These ceramics were exported to large areas of the north-western part of the Roman empire. Before the start of the project, very little information was available on the location of these sites, so it was decided to carry out archaeological prospection (especially field survey) on the basis of a predictive model. This was a fortunate decision, as the predictive model could be used to identify the zones that were insufficiently mapped. During the course of the project it became ever more evident that the location of kiln sites is primarily related to the presence of outcrops of well-suited pottery clay in the vicinity of water courses. These factors could be mapped using geological and topographical maps of the study area. Small imperfections of the model can be attributed to incomplete information in the base maps used. Apart from that, a small portion of the kiln sites is located close to transport routes. As the modelling and the survey were closely connected, a final model with a high predictive power could be produced. The downside of this approach is that it is time-consuming and will lead to high prospection costs.

Chapter 3 describes how in a different area in France (the Valdaine-Tricastin region, located on the east bank of the river Rhône), an attempt was made to demonstrate to what extent field survey will give a biased impression of the location and number of archaeological sites in alluvial zones. While this effect was already known in the Netherlands, in France it was not, and as far as is known this is the only study that has ever tried to quantify it. By using detailed archaeological site registers, an analysis of the location of known archaeological sites was performed on the basis of geological and pedological maps. Separate analyses were carried out for those sites that were exposed at the surface, and for buried sites. A comparison was also made between an analysis that did not take into account the area surveyed, and an analysis based on data from the surveyed zones alone, in order to establish the resulting bias. Furthermore, a rough estimate was made of the number of sites that might be still be present in the area. In all probability, approximately 8,625 undiscovered sites may still be found in the area.

Chapter 4 focuses on the issue of using expert judgment for creating archaeological predictive models. In many cases it is impossible to obtain representative samples of sufficient size to make predictions, and therefore it has become common practice in the Netherlands to create predictive models on the basis of the opinions of experts on the importance of the different site location parameters involved. The main disadvantage of this approach is the difficulty of verifying the experts' judgment, and it is also very hard to combine expert judgment and quantitative methods. In this chapter, the hypothesis is elaborated that Bayesian statistical methods can be helpful in this case, as they are capable of integrating subjective reasoning with statistical sampling theory. However, this can only be done if the experts are able to express their judgment in statistical terms. In other words, they will have to formulate their opinion on the importance of site location parameters, as well as the certainty of their estimates, in a quantitative framework. This is also known as formulating the prior probability distribution. By adding site data from representative samples, it is possible to both test and refine the original estimates. This is known as establishing the posterior probability distribution. In the case study presented in this chapter, the prior distribution was solicited from three experts who were

asked to make a prediction for the municipality of Ede, using multicriteria analysis techniques. This was an effective way to create a prior distribution, but it could also be used to see how much the experts' opinions differed. As there was no opportunity to collect new field data, I decided to use the observations that were already available to see how these would influence the posterior distribution. These data however are not the independent representative sample that one would need in order to perform a Bayesian prediction. It turns out that the correct application of Bayesian statistical methods is complex and needs additional research in the context of archaeological predictive modelling. However, it has the advantage of not only providing a quantitative estimate from expert judgment, but also of giving a margin of error. It is this estimate of uncertainty that is currently lacking in almost every single predictive model made.

CHAPTER 2 The Use of Predictive Modeling for Guiding the Archaeological Survey of Roman Pottery Kilns in the Argonne Region (Northeastern France)[1]

Philip Verhagen and Michiel Gazenbeek[2]

2.1. INTRODUCTION

The Argonne region, situated in the northeast of France (figure 2.1), was an important center of Roman pottery production in north-western Europe. Today it is a quiet area with abundant forest (covering about 50% of the region), but during the First World War it was the stage for fierce frontline fighting between the German and French forces. Numerous trenches are still present in the area, and the remains of ammunition, barbed wire and weapons can be found on many fields.

Figure 2.1. Location of the study area in France

The area is currently experiencing rapid land-use changes that are potentially damaging to the archaeological remains. The most important of these is the conversion of grassland to agricultural land, a development that will increase erosion of the topsoil. Furthermore, new infrastructure in the area is being developed by the French government as part of a revitalization campaign in the area. One of the aims of this campaign is to draw tourists to visit the World War I relics. Also, the new TGV Est high-speed railway, connecting Strasbourg to Paris in 2005-2006, will be running straight through the area.

Given these developments, the *Service Régional d'Archéologie* (SRA) of the region Lorraine decided in 1996 to launch a project[3] to make an inventory of the distribution and state of conservation of pottery kiln sites. The SRA of the region Champagne-Ardennes decided to join the project in 1997, which brought the total study area to 725.62 km^2 in 51 municipalities.

---

[1] This paper also appeared in Mehrer, M. and K. Wescott (eds.), 2006: *GIS and Archaeological Site Location Modeling*. CRC Press, Boca Raton. pp. 447-459.

[2] CNRS, Centre d'Etudes Préhistoire, Antiquité, Moyen Age , Sophia-Antipolis (Valbonne), France. Michiel Gazenbeek contributed the archaeological background for this paper, and coordinated the fieldwork. I wrote the rest of the paper and performed the predictive modelling.

[3] *Les ateliers céramiques gallo-romains d'Argonne: bilan, recherche et gestion patrimoniale*

The aims of the project as expressed by the SRA Lorraine were to:

- establish survey methods appropriate to the region (field walking, geophysical survey and augering);
- decide which sites were to be excavated;
- acquire land and establish archaeological reserves; and
- elaborate protection measures in negotiation with land owners and users.

The survey was to focus on both the known sites as well as sites still undiscovered. As it was clear from the beginning that not all the area could be surveyed, it was necessary to start the project by preparing a predictive map. The survey was carried out in three consecutive years (1996-1998), and before the start of the campaigns a predictive map was prepared to guide the survey; a revised map was produced before the last field campaign, after which the final map was produced and presented to the SRA (Exaltus *et al.*, 1998). The survey consisted of field walking in order to discover kiln sites, augering to establish their extent and precise location, and in selected cases, geophysical survey and excavations to obtain an impression of the remains still present underground. The field walking survey and excavations were carried out by students and staff from the Université de Paris I Panthéon-Sorbonne, the geophysical survey, augering and predictive mapping was performed by RAAP Archeologisch Adviesbureau.

2.2. ARCHAEOLOGICAL CONTEXT

The Argonne area has exported industrial quantities of ceramics all over northwestern Europe during nearly all of Antiquity, from the 1st century AD until at least the 5th century. These products, especially the fine slip ware, are very important for the dating of consumer sites. However, of the production centers themselves, their range of products, their production techniques, and their life span, little was known, even though research had been going on for more than a century. The data collected in the Argonne project helped significantly to better understand the economical and environmental background that ensured the success of the Argonne pottery production, and to understand its place in the regional occupation network.

The Roman settlement pattern of the Argonne area appears as a patchwork of villages surrounded by extensive workshop areas (including pottery kilns), of isolated ceramic and glass production centers (that often mask the associated dwellings), and of *villae* and modest farms spread out over the countryside. Except for some large production centers that were active during the whole period, the workshops were mostly short lived, and moved rapidly from one place to another. This model is evident as early as the 1st century AD when the initial Belgic wares were produced. From this period onward, the geographical expansion of the pottery kilns shows only minor changes until well into the 4th century, even though the products themselves changed completely, moving from Belgic wares to red slip wares and covering a whole range of black slip beakers and common coarse wares as well.

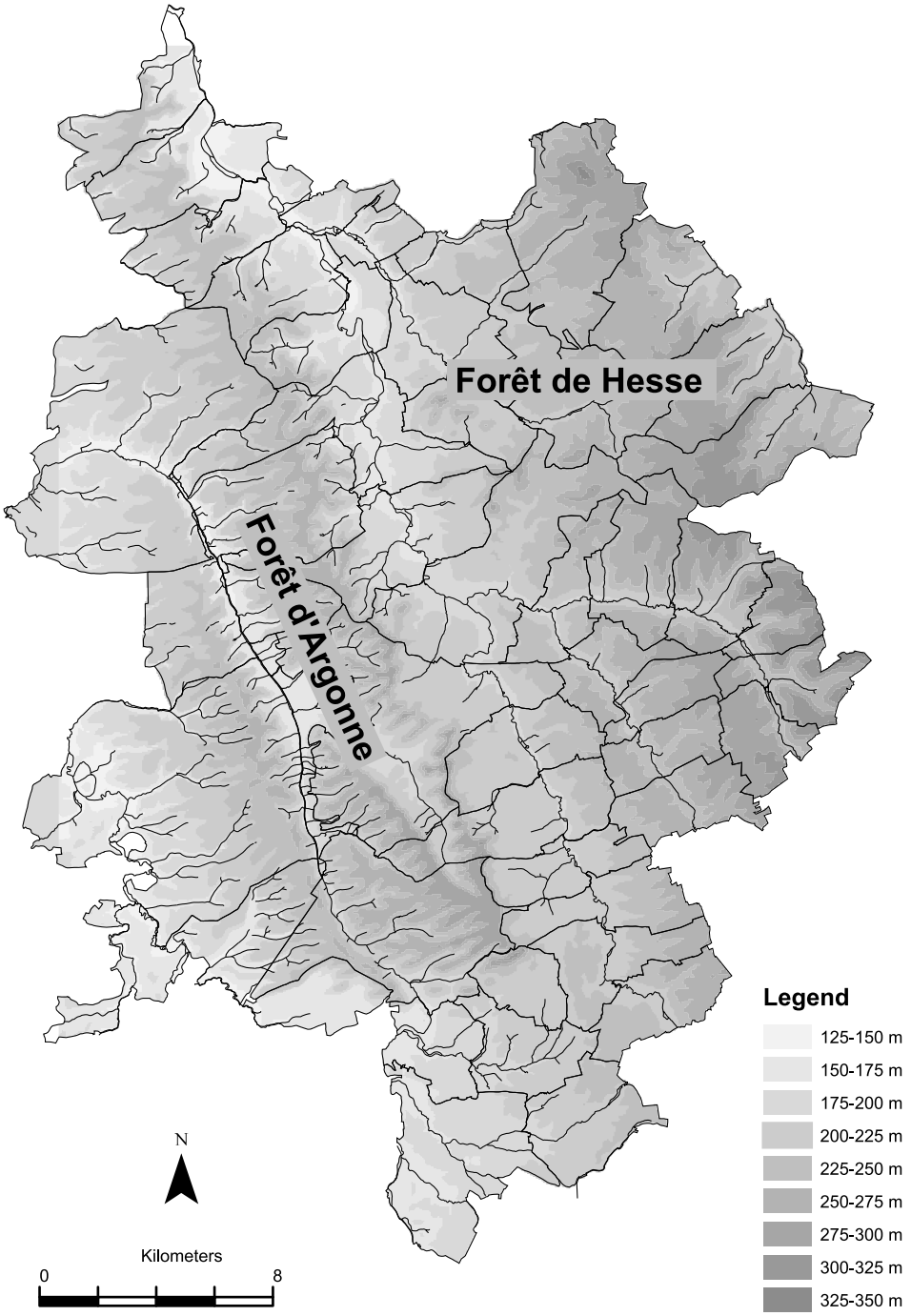


Figure 2.2. Topographical map of the Argonne study area. Source: Institut Géographique National.

The success of the Argonne wares has often been explained by the privileged position of the region close to several major Roman roads that connected it to the main communication network of this part of the Roman Empire, putting it at a relatively short distance of several of the larger cities of northern Gaul, such as Reims and Metz, and providing many markets and redistribution centers for its products. Export over the Marne and Meuse rivers certainly also was very important for the spread of the ceramics, and the nearby town of Verdun on the Meuse probably played a key role in the river transport. However, the importance of production for the regional domestic market has long been underestimated. The fact that the slip wares, the main export product, have been studied in more detail, means that the local markets and the associated potteries have been neglected in research. As it is, coarse wares and roof tiles were produced in large quantities by the various workshops simultaneously with the rest. Our fieldwork also demonstrated the importance of glass and iron workshops in the area. Especially the 3rd and 4th centuries appear to be an important period for the local glass industry, with rows of deep shafts dug into the sandstone that was used as raw material. Its products, such as glass cubes for mosaics, were strongly oriented towards export.

These activities show that the ceramics were far from being the only product manufactured in the region. All together, they testify of a flourishing economy during most of Antiquity, with a very active industry maintaining itself over a long period of time. The importance in this of the natural factors that are characteristic of the Argonne, cannot be underestimated: the geological context is responsible for the environmental conditions that could supply at the same time the basic raw materials needed for different products, and the fuel (wood) necessary to produce the (semi-)finished products. The large area covered by the geological formations involved guaranteed the long life span and the profitability of these industries on such a large scale. The road network that crosses the region, allowing access to the markets, is therefore only a factor that made this particular economical development easier, but it did not command it.

2.3. AREA DESCRIPTION

The Argonne region is an area of undulating hills that form part of the French Ardennes. The area has a general slope towards the northwest, and is dissected by several watercourses. Elevation generally ranges between 175 and 300 m above sea level. The Aire river divides the area in two distinct parts: the Forêt d'Argonne to the west, and the Forêt de Hesse to the east. The World War I frontline runs from west to east through the area (figure 2.2). The geology of the area consists mainly of Jurassic and Cretaceous sedimentary rocks. The oldest formations are found in the east, the youngest in the west of the area. The most important rock types to be found in the area, from old to young, are:

- Kimmeridgian clays and marls;
- Portlandian limestones (*Calcaires du Barrois*);
- Lower Albian greenish clayey sands (*sables verts*);
- Upper Albian yellowish-brown sandy clays (*Argiles du Gault*);
- Cenomanian calcaric sandstones (*Gaize*).

It is assumed that the pottery in the area has been produced using both the *sables verts* and *Argiles du Gault*; however, no definite answer to that question has been found.

Apart from the Jurassic/Cretaceous deposits, colluvial and alluvial deposits are found in the valleys. Along the valleys of the larger watercourses (Aire, Biesme, and Aisne), older alluvial deposits can be found in terraces located 10 to 15 meters above the current valley bottom (figure 2.3).

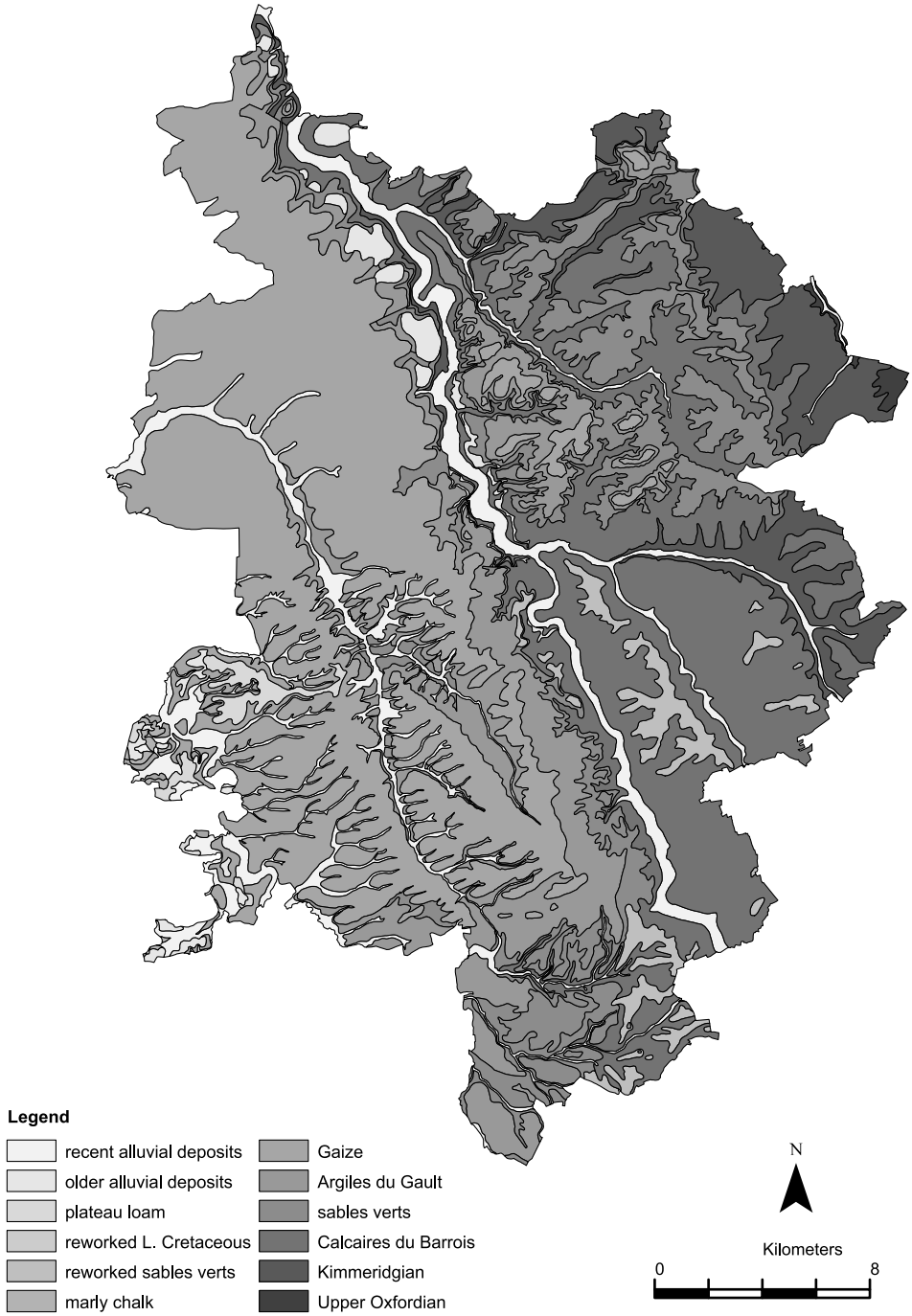


Figure 2.3. Geological map of the Argonne study area. Source: Bureau des Recherches Géologiques et Minières.

The geomorphology of the Forêt d'Argonne is dominated by the *Gaize* sandstones and consists of a strongly dissected plateau with numerous springs where the *Gaize* rests on the *Argiles du Gault*. The Forêt de Hesse is more varied, with *Calcaires du Barrois* in the valleys, followed by *sables verts* and *Argiles du Gault* uphill, and often capped by a *Gaize* sandstone butte. Slopes are more gentle in this area. Springs are not as numerous and can also be found on the transition between *sables verts* and *Calcaires du Barrois*. To the south of the river Vadelaincourt, this geomorphology is replaced by a plateau of *Calcaires du Barrois*. This means that the area where pottery clay can be found is primarily concentrated in the Forêt de Hesse.

Soils in the area are not very well developed. The *sables verts* will weather to a yellowish-brown sandy clay that is sometimes difficult to distinguish from the *Argiles du Gault*. The *Calcaires du Barrois* weather to a brown clay, but it may be strongly eroded on steep slopes. Weathered *Gaize* has not been found in the area, the sandstones are usually covered by a shallow, dark brown, sandy topsoil. Soil erosion is an important phenomenon in the area. The augering campaigns revealed downhill accumulations of colluvial deposits in many places, and often these could be dated to post-Roman times. Especially the *sables verts* and *Argiles du Gault* are highly susceptible to erosion. Currently, however, the rate of erosion is not very high, as most of the area is protected by a vegetative cover of grassland and forest (Timmerman *et al.*, 1998).

2.4. THE FIRST PREDICTIVE MODEL

When confronted with the question of where pottery kiln sites might be located, it is not difficult to understand that the selection of sites for pottery production must have depended on four principal factors:

- proximity to pottery clay;
- proximity to water;
- proximity to fire wood; and
- proximity to transport routes.

The source materials needed (clay, wood, and water) could, in theory, be transported to a different place, but it is not difficult to understand that the pottery is more readily transportable. It is therefore assumed that the proximity to existing transport routes is not a major limiting factor for site placement, whereas the availability of the source materials is. Of these source materials, the location of wood in the Roman period cannot be reconstructed with any accuracy, whereas it can be assumed that the position of watercourses has not changed very much, and the geological formations will still be in the same place. A deductive model of kiln site location will therefore have to start from the assumption that distance to water and pottery clay are the primary site-placement parameters that can be operationalized.

The first results of the fieldwork seemed to confirm this assumption. The kiln sites found were usually located near valley bottoms or springs where *sables verts* and *Argiles du Gault* were available.

From a scientific point of view, the model building should probably have started by preparing a deductive model. The first model built, however, was an inferential one (Gazenbeek *et al.*, 1996). As in many inferential modeling exercises before, the available environmental information (elevation, slope, aspect, geology, and distances to watercourses and springs) was subjected to a χ^2 test on the basis of a small and biased data set. However, the χ^2 test was not used to select the significant variables; all available map layers were reclassified according to site density, and then averaged. Even though the reliability of this model was questionable, it did serve to highlight the weaknesses of the existing archaeological data set.

Until the start of the Argonne Project, the knowledge of the distribution and state of pottery kiln sites in the region was scant. A total of 30 kiln sites had been reported to the French national archaeological database DRACAR, almost exclusively found in the Forêt de Hesse area. These sites were used to construct the first model. Even with this small number of sites, it was clear that the *sables verts* were very important for kiln site location. It also suggested that the known data set was biased, as very few sites were reported in forested areas. This provided two strong objectives for the survey campaigns: to increase the number of known kiln sites to a level where statistical analysis could be done in a meaningful way, and to improve the representativeness of the known site sample.

2.5. THE SECOND PREDICTIVE MODEL

The three field campaigns carried out in 1996-1997 resulted in a dramatic increase of known kiln site locations (table 2.1). Furthermore, the previously reported 30 sites were revisited and checked. In most cases the registered coordinates were wrong, and some sites were withdrawn from the database, either because no traces of the site could be found, or because the site had erroneously been interpreted as a Roman pottery kiln. In September 1997, the number of sites inside the surveyed area had become large enough to justify the construction of a new inferential model of kiln site location. As it was assumed that kiln site location is related to the proximity of key geological formations (notably the *sables verts*), distances to these formations were calculated and subjected to a χ^2 test. The distances were calculated as distances to the geological formation boundary, both outside and inside the geological formation. Distances inside the formation were given a negative value. Furthermore, distances to permanent water courses and springs were used.

Because of the limitations of the χ^2 test, it was necessary to reduce the number of distance categories in such a manner that the number of expected sites per map category should not fall below 5 (see e.g. Thomas, 1976). With a total of 56 sites available inside the surveyed area, this meant that preferably no single map category should be smaller than 8.9% of the area. Therefore, the distance categories used are not equal-interval zones, but equal-area zones, that can be produced in ARC/INFO GRID by using the SLICE command. The resulting distance maps were analyzed for autocorrelation. It turned out that distance to *Argiles du Gault* is rather strongly correlated to distance to *sables verts* and to *Gaize* ($r = 0.59$ and $r = 0.56$ respectively). This can be explained by the fact that *Argiles du Gault* are usually found as a narrow band between *Gaize* and *sables verts*. Furthermore, the distance to *Calcaires du Barrois* is negatively correlated to the distance to *Gaize* ($r = -0.51$). This is due to the fact that the *Gaize* is predominantly found in the west, and *Calcaires du Barrois* in the east of the area. Other, weaker correlations were found between springs and permanent water courses ($r = 0.49$), and between permanent water courses and recent alluvial deposits ($r = 0.44$).

After χ^2 analysis of all variables, the model was based on those variables that were statistically significant at the 95% probability level and not strongly autocorrelated: slope, distance to *sables verts*, distance to *Gaize* and distance to recent alluvial deposits. Although the initial intention was to use only the data from inside the surveyed area, it turned out that the 'kiln hunt' had been successful in the number of sites found, but biased in terms of the area visited. Especially the mainly forested *Gaize* and steep slopes had been avoided by the field walkers, for understandable reasons. To compensate for this effect, the full data set was used, which meant using a possibly biased data set instead of a certainly biased one. The resulting model showed high probabilities for kiln site location in the Forêt de Hesse and along the valley of the Aire, whereas low probabilities were found in the Forêt d'Argonne. However, some outliers were found close to the major rivers

and the presumed location of Roman roads, possibly indicating a preference for location close to transport routes instead of close to the source materials needed.

survey period	*area surveyed*	*inside study region*	*no. kiln sites*	*in surveyed area*
Nov 96	1037.42	1037.42	42	15
Feb 97	1349.03	1308.34	70	47
Sep 97	1381.00	1114.39	74	56
Mar 98	2750.59	1991.19	91	83
TOTAL (ha):	***6518.04***	***5451.34***		

Table 2.1. Area surveyed during the four consecutive field campaigns. Before the start of the survey, only 30 kiln sites were known in the area.

2.6. THE FINAL MODEL

Although the second map was useful in indicating the important zones for kiln site location, it was less well-suited to predict low-probability zones. A large area was still designated medium probability, and this was primarily a consequence of the survey bias. The last field campaign was therefore dedicated to extending the surveyed area into the *Gaize* and steep-sloped zones. The last campaign included over 50% of *Gaize*. This made it possible to produce a model with optimal reliability, from a statistical point of view (Exaltus *et al.*, 1998; figure 2.4).

Apart from that, it was decided to perform a field check to see if the geological maps used were accurate enough for the predictive modeling. The field check (Timmerman *et al.*, 1998) confirmed that the quality of the geological maps is adequate for most of the area, with two notable exceptions. Firstly, where outcrops of geological formations have a limited extent, they are not always mapped. In one particular case a kiln site was found close to a pocket of *sables verts* that was not depicted on the map, and it can be expected that similar locations exist in the area, especially near valley bottoms. Secondly, south of the Forêt d'Argonne a relatively large area of *sables verts* shown on the geological map was not found at the surface, but only at considerable depth. The high probability assigned to this area on the predictive map is therefore incorrect. The final model was satisfying from the point of view of cultural resource management: the area of medium probability had been substantially reduced, and the high-probability zone showed a relative gain of 55.5% (table 2.2). It should, however, be noted that the model can be this specific because the location of kiln sites considered is primarily linked to a very specific location factor, the availability of pottery clay.

	1996		***1997***		***1998***	
probability	*%area*	*%sites*	*%area*	*%sites*	*%area*	*%sites*
low	27.5	2.8	29.9	0.0	63.2	6.6
intermediate	58.6	61.1	52.9	32.4	22.0	23.1
high	13.9	36.1	16.0	67.6	14.8	70.3

Table 2.2. Comparison of the three predictive models made. In the third model, the gain of the high probability zone is 55.5%, and the area of intermediate probability is considerably smaller than for the previous models.

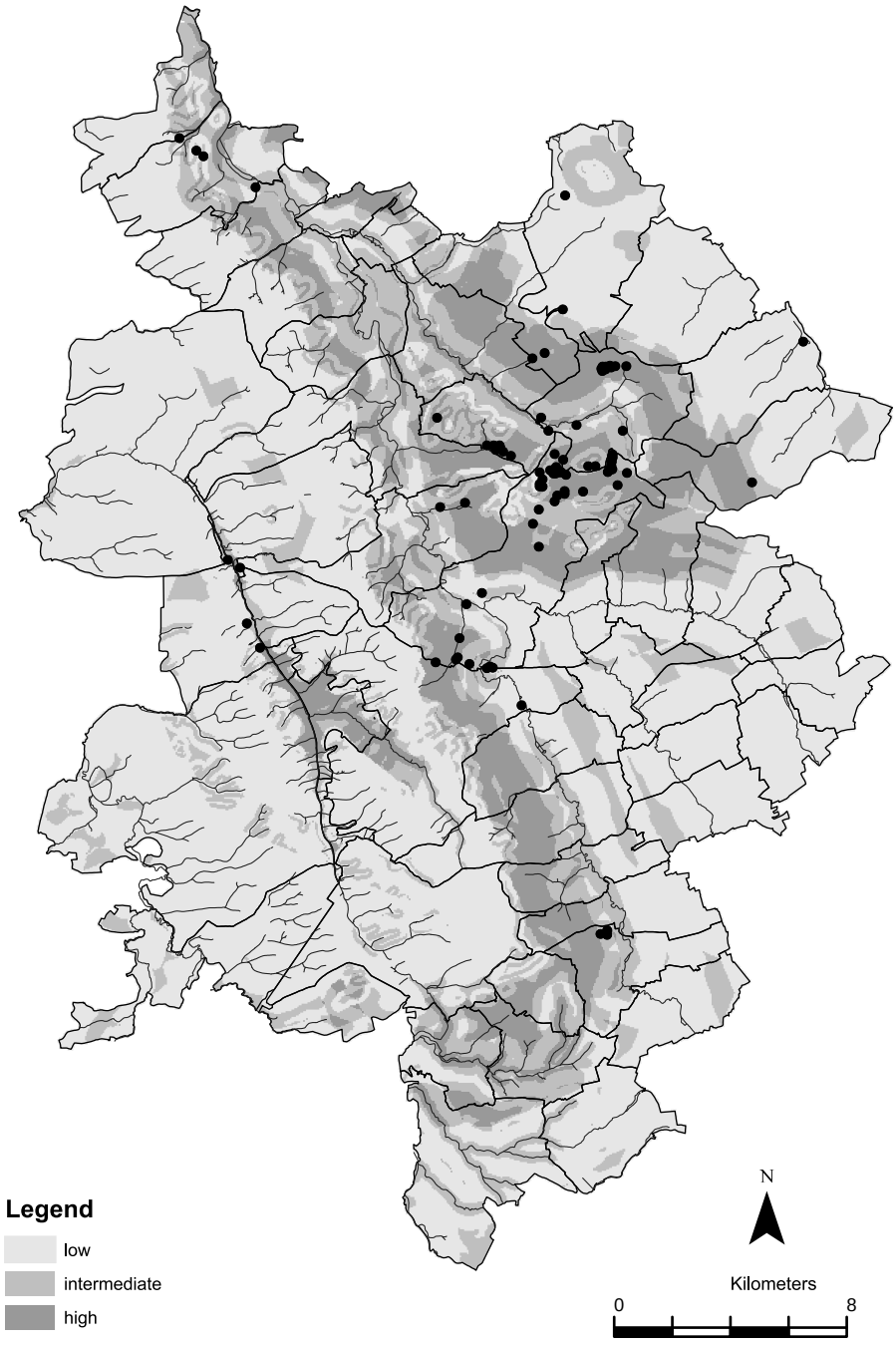


Figure 2.4. The final predictive map. The black dots indicate kiln sites.

2.7. CONCLUSIONS

The Argonne Project succeeded in producing a predictive map that is useful for locating the principal areas of pottery kiln sites with high accuracy. Given the lack of information at the start of the project, this is an impressive achievement.

Nevertheless, it seems that the model can only be this successful because the site type aimed at is highly predictable from a deductive point of view. The availability of pottery clay is the most important location factor to be taken into account, and therefore highly limits the area were kiln sites can be found. In general, this means that predictive models will be most successful when they aim at predicting specific functional site types. This in turn implies that the site sample used for the modeling should be analyzed on specific site types, and the variables used should reflect the possible limiting and attracting factors for locating these sites. This is not a very surprising conclusion, it seems, but one that is frequently overlooked in 'commercial' predictive modeling.

Furthermore, the project made eminently clear that field testing of the predictive maps, both archaeologically as well as geologically, was very useful and, in fact, necessary to obtain a reliable model. It is also clear that the amount of field testing done should be substantial, and it should be guided by the questions that arise from the predictive mapping itself. In practice, this may be a very difficult point to get across to contractors, as predictive modeling is often seen as a means to prevent costly field campaigns. The costs for arriving at the Argonne predictive map were high: a total of 7.5% (54.5 km^2) of the area had to be surveyed in order to obtain the representative site sample needed for the final predictive model. Even with a field campaign primarily looking for sites as easily detectable as pottery kilns, this means a considerable amount of work to be done. In the case of the Argonne Project, four months of field walking survey were done, using 20 students. However, when commercial prices have to be paid, this would represent an investment (at current rates) of approximately 600,000 euros or dollars. In the European public archaeology context, this is a very high price for a field-survey project.

ACKNOWLEDGEMENTS

The authors would like to thank the following people collaborating in the project:

- professor Sander van der Leeuw (Université de Paris I / Panthéon-Sorbonne, UFR03, Histoire de l'Art et Archéologie) and Roel Brandt (former director of RAAP Archeologisch Adviesbureau), coordinators of the Argonne project;
- the colleagues from RAAP Archeologisch Adviesbureau involved in the project, especially Joep Orbons (geophysical survey) and Jan Roymans (field survey); and
- the students from the Vrije Universiteit Amsterdam who did the geological fieldwork: Saskia Gietema, Nicole Rosenbrand and Rinke Timmerman.

REFERENCES

Exaltus, R., J. Orbons, S. Papamarinopoulos, S. van der Leeuw and P. Verhagen, 1998. *Les ateliers céramiques gallo-romains et médiévaux de l'Argonne. Rapport triennal (1996-1998). Volume 3: Carrottages, prospections géophysiques.* Université de Paris I, Paris.

Gazenbeek, M., J. Orbons, T. Spruijt and P. Verhagen, 1996. *Les ateliers céramiques gallo-romains et médiévaux de l'Argonne: bilan, recherche et gestion patrimoniale.* Rapport soumis aux Services Régionaux de l'Archéologie de Lorraine et de Champagne-Ardennes. Université de Paris I, Paris.

Thomas, D.H. 1976. Figuring anthropology. *First principles of probability and statistics.* Holt, Rhinehart & Winston, New York.

R. Timmerman, N. Rosenbrand and S. Gietema, 1998. *Verslag van het bodemkundig/geologisch onderzoek in het kader van het Argonne Project.* Unpublished report. Vrije Universiteit, Amsterdam.

POSTSCRIPT TO CHAPTER 2

The Argonne survey was a unique project for RAAP. It was the only survey that RAAP ever did in France, and it combined traditional field walking, geophysical survey, and core sampling on an unprecedented scale. The project was also remarkable because it used predictive modelling for selecting the areas that needed to be surveyed to improve the model. In terms of increasing the archaeological knowledge of the area, the project was very successful. However, in terms of influencing the attitude of the French archaeological community towards a more positive view of predictive modeling, it failed to be a success. Part of this is due to the fact that publication of the project's results was, up to 2005, restricted to the regional authorities. In September 2000, I presented the predictive modelling methodology on the conference of the European Association of Archaeologists (EAA) in Lisbon. This did not result in publication. The current paper was eventually written in 2002 for publication in Mehrer and Wescott (2006). A separate paper, not specifically focusing on predictive modelling, was published in France in 2003 by Michiel Gazenbeek and Sander van der Leeuw. Michiel Gazenbeek finally presented the project results in September 2004 at the EAA-conference in Lyon, and in October 2004 for a predominantly French audience at the *XXVe Rencontres Internationales d'archéologie et d'histoire d'Antibes*. This has also resulted in a publication (Brandt *et al*., 2005). So, between the end of the project and its final publication for a larger audience, a period of more than 6 years has passed.

RAAP did not succeed in securing more survey work in France. In fact, it was not actively pursued because of the associated logistic and bureaucratic problems. French archaeology is organized in such a way nowadays, that commercial parties are not allowed to participate independently in archaeological research. Furthermore, the Argonne project itself resulted in a substantial financial loss. Given the resistance in France to the application of predictive modelling in archaeological heritage management, it can be doubted whether efforts to promote the methodology of the Argonne project would have been successful. It is not even clear whether the model is actually used for archaeological heritage management in the region at the moment. If so, it is certainly not promoted on the websites of the DRAC of Lorraine and Champagne-Ardennes.

The project also failed to produce any significant spin-off in the Netherlands. This is mainly due to the scale of the project: areas of similar sizes have been studied in the Netherlands, but only for desk-top assessment. Even so, it is clear that the procedure outlined, with reliable survey results being fed back into the model, is essential for further developing predictive models. It is also clear that it is necessary to check the information on geological maps. The scale of the maps used will always influence the outcome of predictive modelling, and it is not surprising that small outcrops of *sables verts* were found that were not shown on the

geological maps. It helps to explain the location of pottery kilns that are not found in the high probability zones, but does not serious influence the basic assumptions of the predictive model at the scale it is made. However, the fact that the maps show *sables verts* at a depth where they could not be exploited compromises the quality of the predictive model, as it is based on the assumption that the deposits were easily accessible for pottery production.

The effort put into the survey clearly resulted in a more accurate and more precise model. Kvamme's gain (which is not mentioned in the paper; see chapter 7) reaches a figure of 0.79 for the high probability zone, which is extremely high. However, the costs for survey were very high, and it is not realistic to expect survey projects of this size being carried out to improve a predictive model in the current Dutch situation.

The model was also about the last inductive/correlative model made by RAAP. Around 1998, RAAP switched its methodology to 'expert judgment' mapping (see also chapter 4). This development, born out of doubts on the applicability of quantitative methods in the Dutch situation, has in retrospect severely thrown back the development of quantitative methods for predictive modelling in the Netherlands. Chapter 7 will go into this issue in more detail.

ADDITIONAL REFERENCES

Brandt, R., M. Gazenbeek, S. van der Leeuw and P. Verhagen, 2005. ‘La gestion du patrimoine archéologique régional ou de l’usage des modèles prédictifs en SIG: l’Argonne, un cas d’école’, in: Berger, J.-F., 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*. Éditions APDCA, Antibes, pp. 93-103.

Gazenbeek, M. and S.E. van der Leeuw, 2003. ‘L'Argonne dans l'Antiquité: étude d'une région productrice de céramique et de verre’. *Gallia* 60:269-317.

Mehrer, M. and K. Wescott (eds.), 2006. *GIS and Archaeological Site Locaton Modeling.* CRC Press, Boca Raton.

CHAPTER 3 The hidden reserve. Predictive modelling of buried archaeological sites in the Tricastin-Valdaine region (Middle Rhône Valley, France) [1]

Philip Verhagen and Jean-François Berger[2]

3.1. INTRODUCTION

The middle Rhône Valley is located at the boundary of the Mediterranean, central European and alpine climate zones. Within the middle Rhône Valley, the Tricastin-Valdaine region is an area of 1086 km^2 located on the east side of the river Rhône (figure 3.1). The area consists of two distinct zones: to the north, the Valdaine Basin comprises the valleys of the Roubion and Jabron rivers, surrounded by Tertiary pre-alpine hills. The most important town in this area is Montélimar. To the south, the Tricastin forms a broad transitional zone between the riverbed of the Rhône, the lateral Holocene alluvial fans and the Tertiary pre-alpine hills. The most important towns in this area are Pierrelatte, St-Paul-Trois-Châteaux and Bollène.

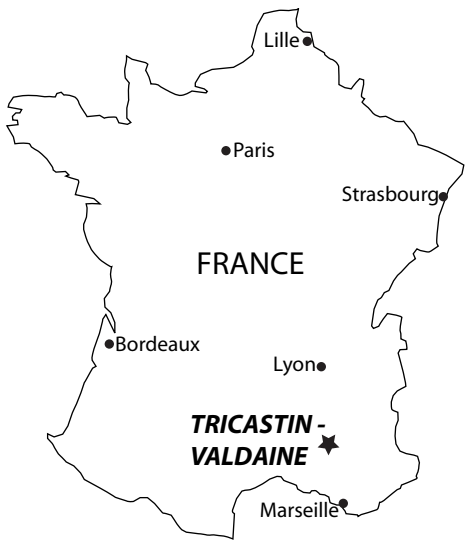


Figure 3.1. Location of the study area in France.

Due to its location at a climatic and geological boundary zone, both vegetation and geomorphological processes in the middle Rhône Valley are highly sensitive to climate change and anthropic impact. The area is known to have a complex history of erosion and sedimentation since the beginning of the Holocene (Brochier *et al.*, 1991; Berger, 1996; Berger *et al.*, 1997; Berger *et al.*, 2000; Berger, 2000; Berger and Brochier, 2000). Because of this landscape dynamic many archaeological remains are known to be buried below the current surface, especially in the alluvial plain of the Rhône. The purpose of this study is to show that the use of the

---

[1] This paper also appeared in Z. Stan?i? and T. Veljanovski (eds.), 2001: *Computing Archaeology for Understanding the Past. Computer Applications and Quantitative Methods in Archaeology CAA2000.* British Archaeological Reports, International Series 931. Archaeopress, Oxford, pp. 219-231. Most of the tables do not appear in this publication, and have been added to this version.

[2] CNRS, Centre d'Etudes Préhistoire, Antiquité, Moyen Age, Sophia-Antipolis (Valbonne), France. Jean-François Berger provided the archaeological and geomorphological background to this paper, the rest of the paper was written by me. The predictive modelling described was very much a joint effort in which I provided the technical and methodological expertise, and Jean-François Berger contributed the (geo-) archaeological data and knowledge.

results of traditional field-walking survey (which will not detect buried sites) for analyzing the relationship between site location and landscape characteristics can lead to both a wrong representation of the distribution of the sites with regard to landscape units, and of actual site quantities. This has been achieved by creating a qualitative predictive map of the area, and by performing a quantitative extrapolation of site densities for the sedimentary areas where most buried sites are found.

3.2. THE PREDICTIVE MODEL

INTRODUCTION

For most predictive modelling studies, the relation of site location to one or more landscape characteristics is inferred by applying an overlay of the known site locations on the cartographic background available. This overlay is then subjected to a quantitative analysis of the observed distribution pattern, an approach also known as inductive modelling (Dalla Bona, 1994). In most cases this analysis is done assuming that the known site sample is representative for the total population. However, this is not necessarily true.

First of all, the method of survey determines which sites will be discovered. It is clear that buried sites will not be detected by means of field walking. However, augering and digging trenches are relatively expensive forms of survey, which are not usually available to amateurs, and even professional archaeologists will not use these forms of survey unless there is a clear necessity. In practice, this means that in most archaeological site databases the number of buried sites will be underestimated.

Furthermore, when the size of the area actually surveyed is not known, it means we do not know where there are no sites, which is equally important for the statistical analysis of site location preference.

Thirdly, the area surveyed (and therefore the site sample) is not usually representative of the total study area. This may be a consequence of difficult access of the terrain, for example because of steep slopes, or because of a research bias for certain areas.

Fortunately, the situation for the Tricastin-Valdaine region is different, as we have both detailed records of buried sites, as well as a mapping of the surveyed zones. However, the current study cannot account for a fourth distorting factor, the differential visibility of archaeological surface remains under different types of land use.

THE TAPHONOMIC MAP: AN INTERPRETATION OF THE LANDSCAPE IN TERMS OF SEDIMENTATION AND EROSION

In order to get a grip on the history of sedimentation and erosion of the area, the landscape has to be interpreted in terms of its geomorphological and pedogenetic history. In order to arrive at a map that could be used as a taphonomic base layer, various geological and soil maps have been digitized, and combined in a GIS. They then have been interpreted in terms of one the following taphonomic categories:

1	Holocene colluvial deposits
2	Pleistocene alluvial fans, stable during the Holocene, with fersiallitic soils
3	Pleistocene alluvial fans, unstable during the Holocene
4	Early / Middle Holocene alluvial fans (Berre and Citelle valleys)
5	Holocene alluvial fans
6	Alluvial plain of the Rhône
7	Early/Middle Holocene terraces
8	Alluvial plains of pre-alpine rivers
9	Stable Pleistocene terraces, with fersiallitic soils
10	Unstable Pleistocene terraces
11	Colluvial basins (*cuvettes*) of the Rhône alluvial plain
12	Colluvial basins (*cuvettes*) of the hinterland
13	Loam and loess formations (Pleistocene)
14	Pleistocene piedmont deposits
15	Weathering resistant rocks (limestones)
16	Rocks with medium resistance to weathering (chalks and marly limestones)
17	Soft rocks (marls, sands, molasses)
18	Holocene alluvio-colluvial deposits

Table 3.1. Basic taphonomic categories for the Tricastin-Valdaine region.

The only base maps available for the whole region are the geological maps 1:50,000 of France. The following sheets have been digitized:

Sheet:	***Name:***	***Publisher:***	***Year:***
842	Crest	BRGM	1976
866	Montélimar	BRGM	1979
XXX-39	Valréas	BRGM	1964
914	Orange	BRGM	1971

These maps have been edited where necessary. The geological map units have then been assigned to one of the 18 taphonomic categories distinguished.

The basic geological information has then been updated with other available information on the geological and pedological conditions in the area. This information comes from three sources:

- a classified remotely sensed image of the Tricastin area (Tounsi et al.,1997);
- a delimitation of the main pedological and sedimentary units obtained during fieldwork in the Valdaine (Berger, 1996) and Tricastin (Berger *et al.*, 1997); and
- existing 1:25,000 pedological maps of the area (Bornand, 1967; 1971).

Essentially, the remotely sensed image gives detailed information on the location of old riverbeds, alluvial fans, terraces and *cuvettes* in the Tricastin. The fieldwork data provide additional information on the location of colluvial and alluvial deposits in the Valdaine (colluvium, alluvial fans, alluvium and *cuvettes*). The pedological maps have been used to find the actual extent of alluvial and colluvial deposits in the Roubion and Jabron valleys, to find the location of stable Pleistocene terraces and alluvial fans in the Valdaine, and the distribution of colluvial deposits and *cuvettes* in the Tricastin. The additional information has been used to update the reclassified geological maps. The final taphonomic map is therefore a patchwork of several maps of varying scale and precision.

Taphonomic categories	Original categories on soil map Valdaine
alluvial plain of the Roubion and Jabron	*sols alluviaux jeunes calcaires (5,6,7)*
Holocene colluvial deposits	*sols colluviaux calcaires (11,12,13,14,18)*
stable Pleistocene terraces	*sols bruns calciques (34,35,36,37,38) des terrasses* *sols bruns et rouges faiblement lessivés (40,41,42,43) des terrasses* *sols (rouges) lessivés (45,46,47) des terrasses*

Table 3.2. Updated units of the taphonomic map based on the soil map of the Bassin Valdainais (Bornand, 1967).

Taphonomic categories	Original categories on soil map Tricastin
colluvial basins (*cuvettes*) of the hinterland	*sols alluviaux calcaires hydromorphes (7,8,9,10)*
colluvial deposits	*complexe de sols de talus de terrasse (18)* *sols régosoliques sableux de bas de pente (25)*

Table 3.3. Updated units of the taphonomic map based on the soil map of the Tricastin (Bornand, 1971).

THE ARCHAEOLOGICAL DATA SET

The archaeological site sample consists of data coming from various sources. Data collected by means of field walking was taken from Beeching *et al*. (1995) and Berger (1996). Data collected by means of digging trenches was taken from Berger (1996), Berger *et al.* (1997; 2000), and from the archaeological surveys and excavations for the TGV Méditerranée (the high-speed railway connection between Lyon and Marseille), which were carried out between 1995 and 1998. Of course the latter method of fieldwork will result in the detection of buried sites, and these data therefore form the most important part of the archaeological database. Data on other sites were collected from literature, and from the archaeological map of the regional archaeological service of the Rhône-Alpes region.

At the regional scale, the relationship between site location dynamics and geomorphological evolution is currently understood in terms of fluvial systems (Berger, 1996; 2000) and can be schematized for the Holocene period according to figure 3.2[3]. In the Rhône hydrological system, the fluvial regime of the rivers changes depending on variations in the ratio of liquid and solid flows. Anastomosed or braided channel systems are unattractive for settlement, except for seasonal activity (Epipalaeolithic, the middle of the early Neolithic, first Iron Age). Flat or convex floodplains, associated with meandering river courses, are more favourable for settlement, but frequent flooding can be a restraint to occupation (late Neolithic, middle Iron Age, Roman high Empire).

[3] note that the study region is actually larger than the fluvial plain of the Rhône, and also takes in some of the transition zone to the Alpine foothills, which is not shown in figure 3.2

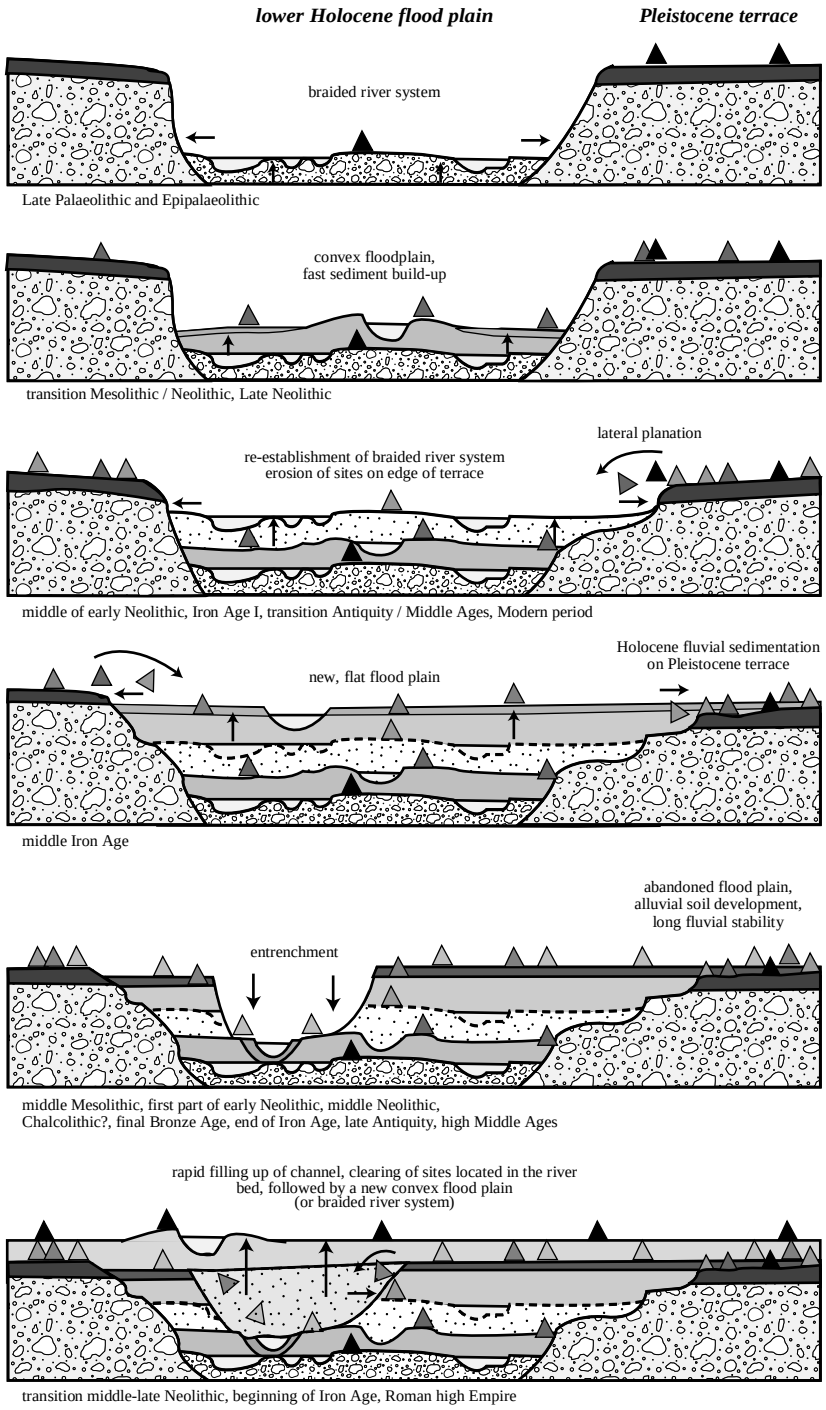


Figure 3.2. Schematic representation of the geomorphological evolution of the Rhône floodplain from the Late Palaeolithic to the Roman period.

The most favourable situation is a meandering river system together with a deep entrenchment of the riverbed, which stabilizes the floodplain for a long period of time (middle to late Mesolithic, middle Neolithic, final Bronze Age, second Iron Age). Evidence is found for a cyclic alternation between stability of watercourses and expansion of occupation in the flood plains on the one hand, and instability of watercourses and depopulation of these same areas on the other hand.

From the sources mentioned, a comprehensive archaeological database was constructed, containing the location of 510 pre- and proto-historic sites or index sites, discovered in different taphonomic contexts (buried in situ, buried in secondary position, at the surface and washed down slopes; see Berger (2000) and Berger *et al.* (2000) for more details). For each site it was documented whether the site was visible at the surface, or if it had been covered by sediment. The term 'site' is not synonymous with find spot in this case. A single find spot may have produced evidence for occupation in more than one period; each of these occupation phases has been stored as a separate site in the database. The occupation phases distinguished have then been regrouped into six different chronological periods (table 3.4). This was done in order to obtain a sufficiently large number of sites per period. The number of sites needed for a reliable analysis of site location in relation to the taphonomic map is approximately 40. For the early Neolithic, this requirement is not met, but the other periods do have sufficient sites to carry out a site location analysis.

period	**visible**	**not visible**	**total**
Epipalaeolithic and Mesolithic	57	8 (12.3%)	**65**
Early Neolithic	21	17 (44.7%)	**38**
Middle Neolithic	57	28 (32.9%)	**85**
Late Neolithic and Chalcolithic	73	43 (37.1%)	**116**
Bronze Age	24	87 (78.4%)	**111**
Iron Age	42	53 (55.8%)	**95**
total	*274*	*236 (46.3%)*	***510***

Table 3.4. Distribution of archaeological sites over 6 chronological periods, and the proportions of visible and non-visible sites.

THE SITE SAMPLE: DEALING WITH THE PROBLEM OF REPRESENTATIVITY

In order to see if non-random sampling influences the site location analysis of the study area, the available site sample was divided into a visible sample and a full sample that also included the non-visible sites. These samples for each period have then been analysed using two geographical analysis windows: the full study region and the area surveyed. In the case of the visible sites, this only encompassed the field walked zones. In the case of the full sample, both the trenched and field walked zones were included (table 3.5). Because of the small number of sites involved, it was not possible to carry out a separate analysis for the trenched zones alone. However, the chances of finding a non-visible site in the trenched zone are much larger than in zones that have only been field walked. As the trenched zone only constitutes a relatively small portion of the total surveyed zone, the actual importance of non-visible sites may be larger than is suggested by the predictive modelling.

window	E/M	EN	MN	LN	BA	IA	Total
FULL(VIS)	57	21	57	73	24	42	***274***
SURV(VIS)	46	12	35	41	9	14	***157***
FULL(ALL)	65	38	85	116	103	95	***502***
SURV(ALL)	51	23	53	60	45	43	***275***

Table 3.5. Distribution of the archaeological sites over the analysis windows. Underlined figures indicate situations where the results of the analysis will be unreliable (n < 40). Windows: FULL(VIS) – whole study region, visible sites; SURV(VIS) – surveyed zones, visible sites; FULL(ALL) – whole study region, all sites; SURV(ALL) – surveyed and trenched zones, all sites. Periods: E/M – Epipalaeolithic/Mesolithic; EN – early Neolithic; MN – middle Neolithic; LN – late Neolithic; BA – Bronze Age; IA – Iron Age**.**

3.3. THE PREDICTIVE MODEL: METHODS APPLIED

In order to analyse the relationships between archaeological site location and the taphonomic map, three separate analyses were undertaken:

χ^2 TEST

A χ^2 test is often used as a first step to see if any statistically significant patterns between site location and map units can be observed. The method has first been suggested by Hodder and Orton (1976), and has been applied on a number of occasions in the Netherlands for predictive modelling purposes (Verhagen, 1995). However, χ^2 in itself does not say anything about the relative importance of map units for site location, and its application as the only statistical tool for predictive modelling has therefore been criticised on a number of occasions (Wansleeben and Verhart, 1992; van Leusen, 1996; Kamermans and Rensink, 1999).

In order to better comply with the limitations of the χ^2 test (the demand of having at least 5 expected sites per map category, which is in turn dependent on the size of the site sample; see e.g. Thomas 1976) the taphonomic map was reclassified into 9 categories (table 3.6). Even so, in some cases the statistical requirements could not be met. In these cases, Yates' correction has been applied to calculate χ^2. It should however be pointed out that in the case of less than 40 observations, the application of χ^2, even with Yates' correction, should be regarded with suspicion.

1	colluvial deposits (1,14,18)
2	stable Pleistocene alluvial fans and terraces (2,9)
3	unstable Pleistocene alluvial fans and terraces (3,10)
4	recent alluvial fans, terraces and riverbeds (4,5,6,7,8)
5	*cuvettes* (11,12)
6	loess formations (13)
7	resistant rocks (15)
8	intermediate rocks (16)
9	soft rocks (17)

Table 3.6. Reclassification of the taphonomic map into 9 categorie

No.	Taphonomic unit	km²	p_a	n	p_s	$K_{j(MAX)}$	rank	$p_{s(CUM)}$	$p_{a(CUM)}$	gain
1	*colluvial deposits*	6.6436	13.5%	10	0.1961	0.4663	3	64.7%	31.1%	33.6%
2	*stable Pleistocene alluvial fans and terraces*	10.4040	21.1%	5	0.0980	0.3828	8	98.0%	83.1%	14.9%
3	*unstable Pleistocene alluvial fans and terraces*	6.6368	13.4%	16	0.3137	0.2372	1	31.4%	13.4%	17.9%
4	*recent alluvial fans, terraces and riverbeds*	7.8804	16.0%	3	0.0588	0.4809	7	90.2%	62.0%	28.2%
5	*cuvettes*	8.3500	16.9%	1	0.0196	0.3187	9	100.0%	100.0%	0.0%
6	*loess formations*	1.7188	3.5%	6	0.1176	**0.5660**	4	76.5%	34.6%	**41.9%**
7	*resistant rocks*	3.1048	6.3%	2	0.0392	0.5637	5	80.4%	40.9%	39.5%
8	*intermediate rocks*	2.5740	5.2%	1	0.0196	0.5466	6	82.4%	46.1%	36.3%
9	*soft rocks*	2.0788	4.2%	7	0.1373	0.3519	2	45.1%	17.6%	27.5%
	TOTAL	**49.3912**	**100.0%**	**51**	**1.00**					

Table 3.7. Example of the calculation of K_j for the Epipalaeolithic and Mesolithic, using the surveyed and trenched zones with the full site sample. n = number of observed sites

RATIO OF SITE TO AREA PROPORTIONS

The ratio of site (p_s) to area (p_a) proportions is a simple and straightforward way to look at the importance of certain map categories for site location. This ratio has for example been used in the Netherlands to create the Indicative Map of Archaeological Values (Deeben *et al*., 1997). However, it does not provide a relative weighting of the categories according to size. This problem is best illustrated by taking the zero site case: a large unit without sites will be less important for site location than a small unit without sites (in order words, it is statistically more significant). Calculated p_s/p_a values however, will give a value of 0 for both units, thereby attributing them equal importance.

In order to account for this effect, Atwell and Fletcher (1985; 1987) suggested calculating a statistic that is described as a relative weight factor for each map unit. In the case of three map-units α, βand γ, the following weights are calculated:

A = a'bc / (a'bc + ab'c + abc')
B = ab'c / (a'bc + ab'c + abc')
C = abc' / (a'bc + ab'c + abc')

where

a, b, c = area proportion of map units α, βand γ;
a', b', c' = site proportion of map units α, βand γ.

This is arithmetically equivalent to dividing each p_s/p_a value found by the sum of all p_s/p_a values, from which it follows that the relative weights calculated with the Atwell-Fletcher method are only normalised p_s/p_a calculations. They will therefore not fully solve the problem of relative weights.

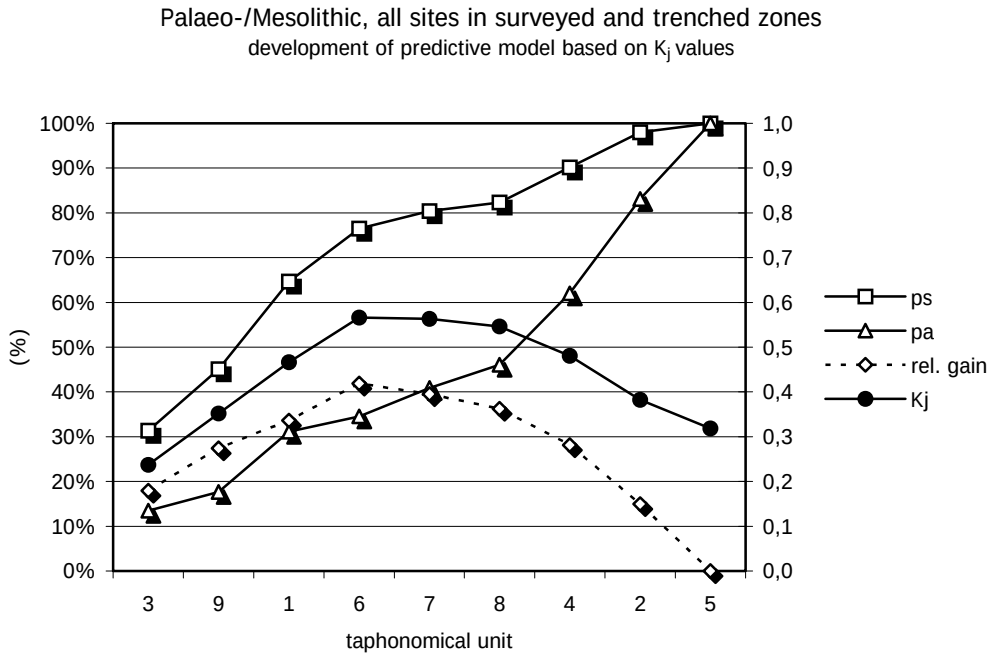


Figure 3.3. Example of the development of the K_j model for the Palaeo-/Mesolithic period, all sites in surveyed and trenched zones. Units 3, 9, 1 and 6 have a strong positive predictive power, units 7 and 8 are more or less neutral, and units 4, 2 and 5 have a strong negative predictive power.

Apart from that, the p_s/p_a calculations do not say anything about the statistical significance of the observed pattern. Atwell and Fletcher (1985; 1987) suggest to test the significance of the pattern by means of comparing the weights to those obtained by simulated site location patterns, a method applied by Wansleeben and Verhart (1992) and Kamermans and Rensink (1999). This analysis depends on the creation of random site distribution maps against which to test the actual pattern. Unfortunately, a random point generating routine is not supplied with ARC/INFO (which was used for the Tricastin-Valdaine model), and time did permit us to write a separate routine, so the simulation was not performed for this study.

K_j METHOD

A more complex method of assessing the importance of map categories for site location is the use of the K_j parameter. This measure was developed by Wansleeben and Verhart (1992), and is defined as follows:

$$K_j = \sqrt{[p_s * (p_s - p_a)]}$$

In the original equation, $p_s - p_a$ is divided by p_w (the proportion of the area without sites); however, this modification is only useful when (hypothetical) site surfaces are used, which is not the case for this model. K_j is calculated for each map category. The category that yields the highest K_j value is considered most

successful. In an iterative procedure K_j is calculated again, including the next most successful category in the model until all categories have been included. The utility of K_j is that it takes into account the relative importance of the observed site densities: a small unit with high site densities will not necessarily be considered the most successful. Each time K_j is calculated, the relative gain (p_s - p_a) can be calculated to assess the performance of the model. A model with high predictive power will have high gain values (Kvamme, 1989). Wansleeben and Verhart (1992) state that the actual performance of the model increases as long the value of K_j increases on each consecutive run. However, we find in a number of instances that the gain is dropping while K_j is still increasing. This is because the equation attributes a higher weight to categories that contain a large number of sites. A gain of 40% can be achieved by a model that contains 50% of the sites on 10% of the surface, but also by a model that contains 80% of the sites on 40% of the surface. The K_j method decides that in the latter case the model performs better, although the gains obtained are equal. However, for the purposes of archaeological resources management, it seems that a model based on gain values is more useful, as the total surface to be considered is smaller.

In the cases that were analysed for this study, we often see categories that contribute strongly to the increase in gain and K_j. Other units will only have a limited effect on the total gain, and there are also units that will strongly decrease the gain of the model. In other terms, these groups may be said to have a positive, neutral and negative predictive power (or high, intermediate and low to use a more traditional terminology). These groups may easily be identified by plotting the development of the model in a graph, and one example of these is given in figure 3.3.

3.4. THE PREDICTIVE MODEL: RESULTS OF SITE LOCATION ANALYSIS

EPIPALAEOLITHIC AND MESOLITHIC (ca. 12,000-6,800 BP; Azilian, Sauveterrian and Castelnovian cultures)

No.	*Taphonomic unit*	*FULL(VIS)*	*SURV(VIS)*	*FULL(ALL)*	*SURV(ALL)*
1	*colluvial deposits*	2.28	1.89	2.00	1.46
2	*stable Pleistocene alluvial fans and terraces*	0.75	0.40	0.82	0.47
3	*unstable Pleistocene alluvial fans and terraces*	4.27	2.09	4.16	2.33
4	*recent alluvial fans, terraces and riverbeds*	0.07	0.14	0.20	0.37
5	*cuvettes*	0.30	0.13	0.26	0.12
6	*loess formations*	**6.14**	**3.68**	**5.39**	**3.38**
7	*resistant rocks*	0.70	0.65	0.62	0.62
8	*intermediate rocks*	0.09	0.39	0.08	0.38
9	*soft rocks*	1.11	3.46	1.39	3.26
	χ^2	**98.73**	**48.74**	**96.42**	**46.42**

Table 3.8. Ratio of site to area proportions for the Epipalaeolithic and Mesolithic. Test value of χ^2 for 99.9% probability level and 8 degrees of freedom is 26.125.

No.	*Taphonomic unit*	*FULL(VIS)*	*SURV(VIS)*	*FULL(ALL)*	*SURV(ALL)*
1	colluvial deposits	2	2	2	3
2	stable Pleistocene alluvial fans and terraces	5	7	5	8
3	unstable Pleistocene alluvial fans and terraces	1	1	1	1
4	recent alluvial fans and riverbeds	9	8	9	7
5	*cuvettes*	7	9	7	9
6	loess formations	3	4	4	4
7	resistant rocks	6	5	6	5
8	intermediate rocks	8	6	8	6
9	soft rocks	4	3	3	2
	$k_{j(max)}$	**0.6429**	**0.6112**	**0.6228**	**0.5660**
	maximum gain	**49.0%**	**46.4%**	**46.7%**	**41.9%**

Table 3.9. Ranking of taphonomic units in the K_j model for the Epipalaeolithic/Mesolithic period, with associated maximum K_j and gain values.

The calculated χ^2 values (with Yates' correction) for the Epipalaeolithic and Mesolithic sites indicate that the taphonomic units are significant for site location at the 99.9% probability level in all analysis windows (table 3.8).

When looking at the ratio of site to area proportions, it is clear that the loess formations show very high p_s/p_a values for all analysis windows. For the full study region, the unstable Pleistocene alluvial fans and terraces also have very high p_s/p_a values. However, in the surveyed zones the soft rocks seem more important. Low p_s/p_a values are observed for the recent alluvial fans, terraces and riverbeds, the intermediate rocks and the *cuvettes*. The recent alluvial fans, terraces and riverbeds are clearly less important for the visible sample than for the full sample. Even though only six buried sites have been observed for this period, they do seem to a have a (limited) effect on the analysis results.

The K_j model that was developed for all analysis windows is rather strong (table 3.9). Maximum K_j and gain values decrease when the surveyed zones are used instead of the full region, and the overall performance is weaker when the full sample is included. It is absolutely clear that map units 3, 1, 9 and 4 are the most important ones for site location. The models are less clear about the units with negative predictive power, so most shifts in ranking are observed for these categories. The pattern of site location observed largely conforms to the pattern obtained with p_s/p_a values, with one notable exception: the position of the loess formations is much less dominant than could be expected from the p_s/p_a calculations. This is because the actual gain obtained by including the small unit of loess formations first in the model is less than the gain that can be achieved by including the large unit of unstable Pleistocene alluvial fans and terraces. This clearly demonstrates that the K_j model is able to perform a relative weighting of map categories.

It can be concluded that the known site sample is representative for the area. Neither the restricted analysis, nor inclusion of the non-visible sites leads to drastic changes in observed site location preference. The most important units for site location are the unstable Pleistocene alluvial fans and terraces, the colluvial deposits, the soft rocks and the loess formations. The observed pattern is distinct, as is demonstrated by the high maximum K_j and gain values observed.

It seems that Epipalaeolithic/Mesolithic settlement is strongly concentrated on the intermediate elevations (with the exception of the stable Pleistocene alluvial fans and terraces), avoiding both the humid zones and the hills. This under-representation of settlements in landscape units that are marked by numerous geomorphological events since the end of the Late Glacial is probably the consequence of taphonomic bias.

The observed absence of buried sites can be attributed to strong erosion of the recent alluvial fans, riverbeds and the lower reaches of the *cuvettes* between 6,400-6,200 BP, associated with the first evidence of agropastoral activity in the south of France and an abrupt hydroclimatic event (Berger, 1996; Berger and Brochier, in press). Furthermore, the Epipalaeolithic/Mesolithic sites found are usually small in size and are characterised by a dispersed lithic scatter, and as such are difficult to detect, even by means of trenching. A geographical bias of Epipalaeolithic/Mesolithic sites is observed for the Valdaine basin. This bias may be the result of selective surveying.

EARLY NEOLITHIC (ca. 6,5000 – 5,800 BP; Cardial and Epicardial cultures, and the transition of Cardial to Chassey culture)

No.	*Taphonomic unit*	*FULL(VIS)*	*SURV(VIS)*	*FULL(ALL)*	*SURV(ALL)*
1	colluvial deposits	2.47	**2.89**	1.59	1.62
2	stable Pleistocene alluvial fans and terraces	2.03	0.38	1.12	0.21
3	unstable Pleistocene alluvial fans and terraces	2.57	1.72	2.49	1.94
4	recent alluvial fans, terraces and riverbeds	0.20	0.54	1.34	**2.18**
5	*cuvettes*	1.61	0.51	0.89	0.26
6	loess formations	**5.56**	2.35	**4.61**	1.25
7	resistant rocks	0.00	0.00	0.00	0.00
8	intermediate rocks	0.00	0.00	0.00	0.00
9	soft rocks	0.86	1.89	0.71	1.03
	χ^2	**NA**	**NA**	**NA**	**NA**

Table 3.10. Ratio of site to area proportions for the early Neolithic. NA = not applicable.

No.	*Taphonomic unit*	*FULL(VIS)*	*SURV(VIS)*	*FULL(ALL)*	*SURV(ALL)*
1	colluvial deposits	1	1	3	3
2	stable Pleistocene alluvial fans and terraces	3	9	5	9
3	unstable Pleistocene alluvial fans and terraces	2	2	2	2
4	recent alluvial fans and riverbeds	8	5	1	1
5	*cuvettes*	5	6	6	8
6	loess formations	4	3	4	4
7	resistant rocks	7	7	8	7
8	intermediate rocks	9	8	9	6
9	soft rocks	6	4	7	5
	$k_{j(max)}$	**0.6774**	**0.5543**	**0.5474**	**0.6101**
	maximum gain	**49.7%**	**41.0%**	**33.2%**	**40.8%**

Table 3.11. Ranking of taphonomic units in the K_j model for the early Neolithic period, with associated maximum K_j and gain values. The quality of the model is questionable, because of the low number of sites involved.

The total number of early Neolithic sites is only 38, which means that calculated χ^2 values are not reliable (table 3.10).

When looking at the ratio of site to area proportions, it is clear that the loess formations show very high p_s/p_a values for the full study region. The unstable Pleistocene alluvial fans and terraces and colluvial deposits also have very high p_s/p_a values. Low p_s/p_a values are observed for the resistant and intermediate rocks. However, in the surveyed zones the position of the loess formations is less dominant. When including

the non-visible sites in the sample, the recent alluvial fans, terraces and river beds become much more important, largely at the expense of the loess formations and colluvial deposits. This again illustrates the importance of including these sites in the analysis, even when working with small samples.

The models developed with the K_j method are unstable (table 3.11). Both maximum K_j and gain values are variable. Units 1, 3 and 6 seems to be most important for site location. It is also obvious that by including the non-visible sample, the importance of recent alluvial fans, terraces and riverbeds becomes much larger.

It is difficult to draw conclusions about site location preference for the early Neolithic because of the low number of sites. This low density is in part the consequence of the major erosion phase occurring between 6,400 and 6,200 BP (Berger and Brochier, 2000). The K_j model for the full sample indicates one important change compared to the Epipalaeolithic and Mesolithic: the important position of the recent alluvial fans, terraces and riverbeds. Since only the earliest horizon of early Neolithic occupation is destroyed or reworked, this implies that the more recent horizons have been preserved under younger alluvial deposits.

MIDDLE NEOLITHIC (ca. 5,800 – 5,000 BP; Chassey culture)

No.	*Taphonomic unit*	*FULL(VIS)*	*SURV(VIS)*	*FULL(ALL)*	*SURV(ALL)*
1	colluvial deposits	2.13	1.98	1.73	1.54
2	stable Pleistocene alluvial fans and terraces	1.31	0.52	1.25	0.54
3	unstable Pleistocene alluvial fans and terraces	**4.27**	2.75	**3.34**	**2.53**
4	recent alluvial fans, terraces and riverbeds	0.30	0.37	0.85	1.06
5	*cuvettes*	0.59	0.00	0.79	0.22
6	loess formations	3.07	1.61	2.06	1.08
7	resistant rocks	0.00	0.00	0.12	0.00
8	intermediate rocks	0.09	0.00	0.12	0.00
9	soft rocks	1.27	**3.25**	1.06	2.24
	χ^2	**77.11**	**NA**	**61.80**	**36.44**

Table 3.12. Ratio of site to area proportions for the middle Neolithic. Test value of $\chi 2$ for 99.9% probability level and 8 degrees of freedom is 26.125. NA = not applicable.
is.

No.	*Taphonomic unit*	*FULL(VIS)*	*SURV(VIS)*	*FULL(ALL)*	*SURV(ALL)*
1	colluvial deposits	2	2	2	2
2	stable Pleistocene alluvial fans and terraces	4	5	3	9
3	unstable Pleistocene alluvial fans and terraces	1	1	1	1
4	recent alluvial fans and riverbeds	8	8	5	4
5	*cuvettes*	6	9	7	8
6	loess formations	5	4	6	5
7	resistant rocks	7	7	8	7
8	intermediate rocks	9	6	9	6
9	soft rocks	3	3	4	3
	$k_{j(max)}$	**0.6393**	**0.6360**	**0.4991**	**0.5402**
	maximum gain	**46.6%**	**48.8%**	**28.8%**	**34.4%**

Table 3.13. Ranking of taphonomic units in the K_j model for the middle Neolithic period, with associated maximum K_j and gain values.

The calculated χ^2 values (with Yates' correction where applicable) for the middle Neolithic sites indicate that the taphonomic units are significant for site location at the 99.9% probability level. For the visible sample however, the number of sites drops below 40 for the surveyed zones, which makes the χ^2 calculation unreliable (table 3.12).

For the full study region, the unstable Pleistocene alluvial fans and terraces and the loess formations have the highest site densities. Very low site densities are found on the resistant and intermediate rocks. For the surveyed zones, the most important units are the soft rocks and the unstable Pleistocene alluvial fans and terraces, and low densities are observed for the *cuvettes*, resistant rocks and intermediate rocks. When comparing the visible sample to the full sample, there is marked increase in p_s/p_a for the recent alluvial fans and riverbeds, again pointing to the importance of the non-visible sites for the analysis. The calculation of K_j for the visible sites results in strong models with high maximum K_j and gain values (table 3.13). However, when the non-visible sites are included in the models, they are considerably weaker. In spite of this, the importance of units 1, 3 and 9 is very clear for both the visible and full sample. Although the recent alluvial fans, terraces and riverbeds become more important when looking at the full site sample, the effect is less marked than for the early Neolithic.

It can be concluded that the known visible site sample is representative of the area. However, the inclusion of the non-visible sites shows that there is a strong effect of underestimation of the importance of the recent alluvial fans, terraces and riverbeds. A preference can be observed for the unstable Pleistocene alluvial fans and terraces, the colluvial deposits and soft rocks. Compared to the Epipalaeolithic and Mesolithic however, the recent alluvial fans, terraces and riverbeds are more important, at the expense of the loess formations. As the K_j models developed for the full sample for the middle Neolithic are not very strong, this implies that settlement is more dispersed than during the Epipalaeolithic and Mesolithic, which might be related to a diversification in subsistence strategies[4].

LATE NEOLITHIC (ca. 5,000 – 3,700 BP; including Chalcolithic)

No.	*Taphonomic unit*	*FULL(VIS)*	*SURV(VIS)*	*FULL(ALL)*	*SURV(ALL)*
1	colluvial deposits	1.78	1.69	1.12	0.99
2	stable Pleistocene alluvial fans and terraces	1.02	0.45	0.64	0.32
3	unstable Pleistocene alluvial fans and terraces	2.59	1.84	**2.45**	2.23
4	recent alluvial fans, terraces and riverbeds	0.41	0.16	0.95	1.15
5	*cuvettes*	0.46	0.30	0.44	0.30
6	loess formations	**3.20**	2.75	2.01	1.92
7	resistant rocks	1.10	1.46	1.56	1.06
8	intermediate rocks	0.42	0.87	0.36	0.64
9	soft rocks	1.24	**2.77**	1.09	**2.38**
	χ^2	**34.62**	**25.20**	**36.81**	**28.87**

Table 3.14. Ratio of site to area proportions for the late Neolithic. Test value of $\chi 2$ for 99.9% probability level and 8 degrees of freedom is 26.125.

[4] In fact, this is not a very likely explanation. The introduction of agriculture in the Neolithic is more likely to have led to a preference for specific soil types. It is more probable that any earlier sites in the alluvial zones were eroded.

No.	*Taphonomic unit*	*FULL(VIS)*	*SURV(VIS)*	*FULL(ALL)*	*SURV(ALL)*
1	colluvial deposits	2	2	4	5
2	stable Pleistocene alluvial fans and terraces	6	7	7	9
3	unstable Pleistocene alluvial fans and terraces	1	1	1	1
4	recent alluvial fans and riverbeds	9	9	3	2
5	*cuvettes*	7	8	8	8
6	loess formations	5	4	6	4
7	resistant rocks	4	5	2	6
8	intermediate rocks	8	6	9	7
9	soft rocks	3	3	5	3
	$k_{j(max)}$	**0.4746**	**0.5511**	**0.4025**	**0.4894**
	maximum gain	**28.3%**	**37.3%**	**19.2%**	**28.2%**

Table 3.15. Ranking of taphonomic units in the K_j model for the late Neolithic period, with associated maximum K_j and gain values

The calculated χ^2 values (with Yates' correction where applicable) for the late Neolithic sites indicate that the taphonomic units are significant for site location at the 99.9% probability level, with the exception of the visible sample for the surveyed zones (table 3.14).

The p_s/p_a ratios obtained for the full study region show that the loess formations and the unstable Pleistocene alluvial fans and terraces have the highest site densities. No very low site densities are found. In the restricted zones, the most important units are the soft rocks, the loess formations and the unstable Pleistocene alluvial fans and terraces. When comparing the visible sample to the full sample, there is a marked increase in p_s/p_a for the recent alluvial fans and riverbeds for the surveyed and trenched zone, again pointing to the importance of the non-visible sample for the analysis.

The calculation of K_j for the full study region produces a weaker model than for the surveyed zones (table 3.15). It is clear that unit 3 is the most important unit for site location; however, the models differ considerably in attributing a ranking to most other units. For the full sample, the recent alluvial fans, terraces and riverbeds are clearly more important than for the visible sample. In all cases, the *cuvettes* and stable Pleistocene alluvial fans and terraces have a negative predictive power.

It can be concluded that the known site sample is not representative of the area, as the performance of the K_j models differs considerably when looking at the restricted zones. From the available data it can be deduced that the higher elevations (soft rocks, intermediate rocks and resistant rocks) may have been neglected in previous surveys. Furthermore, the inclusion of the non-visible sites shows that there is a very strong effect of underestimation of the importance of the recent alluvial fans, terraces and riverbeds. A preference can be observed for the recent alluvial fans, terraces and riverbeds, unstable Pleistocene alluvial fans and terraces, and soft rocks. As the K_j models developed for the full sample for the late Neolithic are not very strong, this implies that settlement is rather dispersed. In fact, during this period a slight shift in occupation towards the higher elevations is observed.

BRONZE AGE (ca. 3,700 – 2,700 BP)

No.	*Taphonomic unit*	*FULL(VIS)*	*SURV(VIS)*	*FULL(ALL)*	*SURV(ALL)*
1	colluvial deposits	0.72	0.96	0.76	0.83
2	stable Pleistocene alluvial fans and terraces	0.44	0.00	0.31	0.00
3	unstable Pleistocene alluvial fans and terraces	1.69	2.29	1.97	1.65
4	recent alluvial fans, terraces and riverbeds	0.53	0.73	1.11	**2.23**
5	*cuvettes*	0.70	0.67	1.31	0.92
6	loess formations	0.00	0.00	1.13	0.64
7	resistant rocks	1.67	**3.32**	**2.24**	0.71
8	intermediate rocks	0.22	0.00	0.20	0.43
9	soft rocks	**3.39**	2.53	1.05	1.58
	χ^2	**NA**	**NA**	**41.90**	**25.67**

Table 3.16. Ratio of site to area proportions for the Bronze Age. Test value of $\chi 2$ for 99.9% probability level and 8 degrees of freedom is 26.125. NA = not applicable.

No.	*Taphonomic unit*	*FULL(VIS)*	*SURV(VIS)*	*FULL(ALL)*	*SURV(ALL)*
1	colluvial deposits	4	4	7	5
2	stable Pleistocene alluvial fans and terraces	7	9	8	9
3	unstable Pleistocene alluvial fans and terraces	3	1	3	2
4	recent alluvial fans and riverbeds	8	5	2	1
5	*cuvettes*	5	6	4	4
6	loess formations	6	7	6	7
7	resistant rocks	2	2	1	6
8	intermediate rocks	9	8	9	8
9	soft rocks	1	3	5	3
	$k_{j(max)}$	**0.5130**	**0.5688**	**0.4575**	**0.4974**
	maximum gain	**38.2%**	**41.0%**	**24.8%**	**30.8%**

Table 3.17. Ranking of taphonomic units in the K_j model for the Bronze Age, with associated maximum K_j and gain values. The model is unreliable for the visible site sample because of the low number of sites involved.

The calculated χ^2 values for the Bronze Age sites are not reliable for the visible site sample, as it only includes 24 sites (table 3.16). For the full sample the calculated value (with Yates' correction when applicable) is significant for site location at the 99.9% probability level for the whole region, but for the surveyed zones it is not.

When looking at the p_s/p_a ratios obtained for the visible sample, the soft rocks clearly exhibit the highest values when looking at the whole region. However, when looking at the surveyed zones, the resistant rocks are most important, followed by the soft rocks and unstable Pleistocene alluvial fans and terraces. Low values are found for the loess formations, intermediate rocks and stable Pleistocene alluvial fans and terraces. When including the non-visible sample, the resistant rocks seem most important when looking at the whole region. Within the surveyed zones, the recent alluvial fans, terraces and riverbeds are most important. The observed patterns seem highly irregular; however, when looking at the two reliable samples, it is obvious that a change in importance can be observed from resistant rocks to recent alluvial fans, terraces and riverbeds.

When looking at the models developed with the K_j method, it is clear that the performance of the models is better when only looking at the visible sample (table 3.17). The inclusion of the non-visible sample results in a more important position for the recent alluvial fans, terraces and riverbeds. When looking at the surveyed zones for the full sample, the importance of the resistant rocks for site location is clearly diminished.

Obviously, the visible site sample is wrongly representing both the quantities of Bronze Age sites, as well as their distribution in the landscape. The visible sample contains relatively more sites on resistant rocks. This can be related to a small amount of cave settlements on this unit near Donzère, which are not included in the surveyed zones. This is to due to the history of regional archaeological research, which privileged karstic areas (secondary calcareous formations) until the last decade (Berger *et al.*, 2000). Apart from that, the role of the recent alluvial fans, terraces and riverbeds is clearly underestimated when only looking at the visible sample. When looking at the total sample, a preference is found for site location on recent alluvial fans, terraces and riverbeds, unstable Pleistocene alluvial fans and terraces, and soft rocks. The preference for the river valleys, together with an increasing importance of the *cuvettes* indicates that humid zones become more important for settlement. However, the K_j models developed for the full sample are not very strong. Together with the low χ^2 values found this implies that the settlement pattern is highly dispersed and might even be uniformly distributed. This is certainly due to the strong increase in occupation during the final Bronze Age (3,200-2,700 BP). The original settlement pattern may however have been more strongly concentrated in the alluvial plains. Many Bronze Age sites have been found in secondary position down terrace slopes or in river channels as a consequence of a major erosion phase between 2,700 and 2,300 BP (known as the first Iron Age hydroclimatic crisis), following a long phase of fluvial stability during the Bronze Age period (Berger *et al.*2000). This might imply that many more have been totally destroyed in this period.

IRON AGE (2,700 – 2,200 BP; Hallstatt and La Tène cultures)

No.	*Taphonomic unit*	*FULL(VIS)*	*SURV(VIS)*	*FULL(ALL)*	*(SURV(ALL)*
1	colluvial deposits	0.41	0.00	1.00	1.04
2	stable Pleistocene alluvial fans and terraces	0.51	0.33	0.45	0.33
3	unstable Pleistocene alluvial fans and terraces	1.29	1.96	1.28	1.38
4	recent alluvial fans, terraces and riverbeds	0.30	0.00	1.20	1.75
5	*cuvettes*	2.01	0.43	**1.60**	0.55
6	loess formations	0.00	0.00	1.23	1.34
7	resistant rocks	**2.62**	**6.41**	1.48	**2.22**
8	intermediate rocks	0.61	1.28	0.33	0.45
9	soft rocks	2.15	1.62	1.24	0.55
	χ^2	**29.30**	**NA**	**17.41**	**15.56**

Table 3.18. Ratio of site to area proportions for the Iron Age. Test value of χ2 for 99.9% probability level and 8 degrees of freedom is 26.125. NA = not applicable.

No.	Taphonomic unit	FULL(VIS)	SURV(VIS)	FULL(ALL)	SURV(ALL)
1	colluvial deposits	8	7	6	4
2	stable Pleistocene alluvial fans and terraces	6	8	8	9
3	unstable Pleistocene alluvial fans and terraces	3	2	4	3
4	recent alluvial fans and riverbeds	9	9	1	1
5	*cuvettes*	4	6	5	8
6	loess formations	5	5	7	5
7	resistant rocks	1	1	2	2
8	intermediate rocks	7	4	9	7
9	soft rocks	2	3	3	6
	$k_{j(max)}$	**0.5145**	**0.6834**	**0.4040**	**0.4574**
	maximum gain	**37.1%**	**54.5%**	**18.2%**	**26.5%**

Table 3.19. Ranking of taphonomic units in the Kj model for the Iron Age, with associated maximum Kj and gain values. The model is unreliable for for the visible site sample in the surveyed zones because of the low number of sites involved.

The calculated χ^2 values for the Iron Age sites are not reliable for the visible sample in the surveyed zones, as only 25 sites are found there (table 3.18). In the other cases, the calculated values (with Yates' correction when applicable) indicate that the taphonomic units are not significant for site location at the 99.9% probability level for the whole region, with the exception of the visible site sample for the whole study region.

The p_s/p_a ratios obtained for the visible site sample for the full study region indicate three important units: the resistant rocks, the soft rocks and the *cuvettes*. For the surveyed zones, the resistant rocks are clearly the most important. When looking at the full sample for the whole region, very little difference in site density is found. However, in the restricted zones the resistant rocks show the highest p_s/p_a ratios. It is interesting to observe that the importance of the resistant rocks increases when looking at the surveyed zones: this seems to indicate that more sites (fortified *oppida*) may be found on this unit, but may have been overlooked outside the surveyed areas.

The calculation of K_j results in rather weak models, with the exception of the visible sample for the restricted zones (table 3.19). This better performance for these zones is associated with the clear preference for resistant rocks in these windows. A dramatic shift in importance is observed for the recent alluvial fans, terraces and riverbeds when looking at the full sample for the full region and the surveyed and trenched zones.

The known site sample is not representative for the whole region, as the performance of the K_j models is stronger for the surveyed zones. The large differences in K_j and gain values between the visible and full sample indicate that the visible sample is not representative for the actual settlement distribution. From the full sample it can be concluded that the recent alluvial fans, terraces and riverbeds, resistant rocks and unstable Pleistocene alluvial fans and terraces are most important for site location. The important position of resistant rocks is related to the existence of hill forts. Most of the sites found by trenching under alluvium can be identified as small farm sites dated to the second Iron Age.

However, the weak performance of the K_j models for the full sample together with the low χ^2 values can be taken as an indication that settlement distribution in the Iron Age is close to random with regard to the taphonomic units distinguished. This is totally contradictory to the existing theories on Iron Age site location before the trenching campaigns started (*cf.* Odiot, 1985).

CONCLUSIONS

The results of the site location analysis for the area point to large differences in reliability of the site samples. Especially for those periods where large numbers of sites have been discovered by trenching (notably the early Neolithic, late Neolithic, Bronze Age and Iron Age) it is clear that the sedimentary areas are much more important for site location than can be deduced from the visible site sample alone. Furthermore, the visible site sample is not always representative for the area, as becomes clear for both the late Neolithic and the Iron Age. Any predictive map to be made for the area will therefore have to include both the extent of the prospected zones as well as the information on buried site locations.

It also seems clear that site location characteristics become less pronounced in the later occupation phases. The strong preference for the intermediate elevations in the Epipalaeolithic and Mesolithic is gradually replaced by a rather dispersed settlement pattern in the Bronze Age and Iron Age. This can largely be attributed to the effect of differential conservation for the various periods – however, for purposes of archaeological resources management this is not particularly relevant. From the analysis results it follows that the definition of zones of Epipalaeolithic and Mesolithic settlements will be much easier than for the later periods. It may be possible that for these later periods, site location preferences are be dependent on landscape elements that are not included in the taphonomic map, either because they are too small in size, or because the soil units that were aggregated to create the map are not particularly relevant. A more detailed reconstruction of the (palaeo-)landscape is therefore needed to arrive at an alternative explanation of site location preferences for these periods.

Window	*E/M*	*EN*	*MN*	*LN*	*BA*	*IA*
FULL(VIS)	YES	NA	YES	YES	NA	YES
SURV(VIS)	YES	NA	NA	NO	NA	NA
FULL(ALL)	YES	NA	YES	NO	YES	NO
SURV(ALL)	YES	NA	YES	YES	NO	NO

Table 3.20. Results of $\chi 2$ test for each period and analysis window. YES = significant at 99.9% probability level; NO = not significant at 99.9% probability level; NA = not applicable (number of sites not sufficient for reliable test).

From the point of view of reliability, the best model will be based on the results of the full sample for the surveyed and trenched zones (analysis window SURV(ALL)). The results of the χ^2 test (table 3.20) indicate that with the exception of the early Neolithic, sufficient sites are available in order to construct a model for this window based on taphonomic categories. However, for the Bronze Age and Iron Age significance requirements are not met, so a predictive map based on the χ^2 test for these periods will not be very useful, as the settlement pattern can also be explained by a random distribution of sites. This leaves us with three periods where a useful predictive model based on the χ^2 test can be constructed with a high degree of confidence, the Epipalaeolithic/Mesolithic, middle Neolithic and late Neolithic. These models can then be weighted by means of p_s/p_a ratio, or by using the Atwell-Fletcher method.

It is interesting to observe that valid K_j models may still be developed in cases where the χ^2 test does not meet significance requirements. This is explained by the emphasis placed by the K_j method on the combination of large number of sites and large area units, whereas χ^2 can better be regarded as a measure of concentration of sites. The value of χ^2 depends on the difference between observed and expected sites. Within the total sample, this difference is potentially largest for small area units. In the case where 10 sites are found where only 1 was expected (a difference of 9), the resulting value of χ^2 will be 9.00. If 59 sites are found

where 50 are expected however, χ^2 will only be 1.62. In the K_j model such a unit will nevertheless be considered more important than the smaller unit with only 10 sites (as is clearly observed for Epipalaeolithic and Mesolithic, where the strong concentration of sites on loess formations is not reflected in the K_j model).It remains an open question if K_j models may be used when the number of available sites is very low. The results obtained for the early Neolithic seem to indicate that with low numbers of sites the models become unstable.

A simple method to arrive at a predictive map based on K_j is by plotting the development of the model, either by gain or by the value of K_j itself. Table 3.21 shows the gain per map category for each chronological period. These individual gain values can be used as a weight factor for each map category. If a qualitative mapping is desired, the mapping becomes a question of deciding on the limits between positive, neutral and negative predictive power (table 3.22).

This method of weighting is preferable over the Atwell-Fletcher method. Table 3.23 shows the weights obtained with the Atwell-Fletcher method. The actual ranking of the units obtained with the K_j method is different, reflecting their relative importance and the 'zero site' categories are given different weights, depending on their size. However, the method applied does not say anything about absolute site densities. In order to compare the weights per period, a correction should be applied for the total amount of sites (table 3.24). When these weighed values are plotted in a histogram (figure 3.4) it is immediately clear which units are when important for site presence or absence.

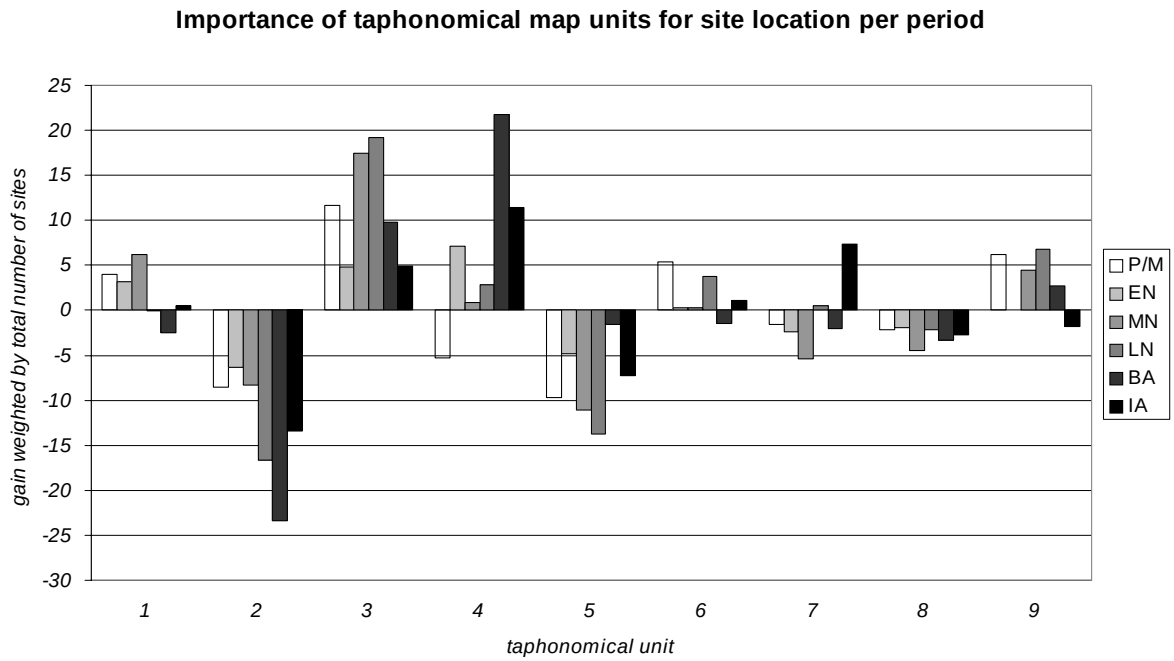


Figure 3.4. Development of the importance of the taphonomic map units for site location through time, by weighting the gain values from the K_j model by the total number of sites involved.

No.	Taphonomic unit	E/M	EN	MN	LN	BA	IA
1	colluvial deposits	6,2%	8,3%	7,3%	-0,1%	-2,3%	0,5%
2	stable Pleistocene alluvial fans and terraces	-13,2%	-16,7%	-9,7%	-14,4%	-21,1%	-14,1%
3	unstable Pleistocene alluvial fans and terraces	17,9%	12,6%	20,5%	16,6%	8,8%	5,2%
4	recent alluvial fans and riverbeds	-8,1%	18,8%	1,0%	2,4%	19,6%	12,0%
5	cuvettes	-14,9%	-12,6%	-13,1%	-11,9%	-1,4%	-7,6%
6	loess formations	8,3%	0,9%	0,3%	3,2%	-1,3%	1,2%
7	resistant rocks	-2,4%	-6,3%	-6,3%	0,4%	-1,8%	7,7%
8	intermediate rocks	-3,3%	-5,2%	-5,2%	-1,9%	-3,0%	-2,9%
9	soft rocks	9,5%	0,1%	5,2%	5,8%	2,5%	-1,9%

Table 3.21. Gain development of K_j models for each chronological period (full site sample, surveyed zones).

No.	Taphonomic unit	E/M	EN	MN	LN	BA	IA
1	colluvial deposits	+	+	+	+/-	+/-	+/-
2	stable Pleistocene alluvial fans and terraces	-	-	-	-	-	-
3	unstable Pleistocene alluvial fans and terraces	+	+	+	+	+	+
4	recent alluvial fans and riverbeds	-	+	+/-	+/-	+	+
5	cuvettes	-	-	-	-	+/-	-
6	loess formations	+	+/-	+/-	+/-	+/-	+/-
7	resistant rocks	+/-	-	-	+/-	+/-	+
8	intermediate rocks	+/-	-	-	+/-	+/-	+/-
9	soft rocks	+	+/-	+	+	+/-	+/-

Table 3.22. Example of a qualitative interpretation of the results obtained with the K_j method. A regrouping has been performed in three categories: positive predictive power (+), neutral predictive power (+/-) and negative predictive power (-).

No.	Taphonomic unit	E/M	EN	MN	LN	BA	IA
1	colluvial deposits	11.8%	19.1%	16.7%	9.0%	9.2%	10.8%
2	stable Pleistocene alluvial fans and terraces	3.8%	2.4%	5.8%	2.9%	0.0%	3.4%
3	unstable Pleistocene alluvial fans and terraces	18.9%	22.9%	27.4%	20.3%	18.4%	14.4%
4	recent alluvial fans and riverbeds	3.0%	25.7%	11.5%	10.5%	24.8%	18.2%
5	*cuvettes*	0.9%	3.0%	2.4%	2.7%	10.2%	5.7%
6	loess formations	27.3%	14.7%	11.8%	17.5%	7.1%	13.9%
7	resistant rocks	5.0%	0.0%	0.0%	9.7%	7.9%	23.1%
8	intermediate rocks	3.0%	0.0%	0.0%	5.8%	4.7%	4.6%
9	soft rocks	26.3%	12.2%	24.3%	21.6%	17.6%	5.8%

Table 3.23. Relative weighting obtained by the Atwell-Fletcher method for each taphonomic unit per period, based on the full site sample for the surveyed and trenched zones.

No.	Taphonomic unit	E/M	EN	MN	LN	BA	IA
1	colluvial deposits	4.0	3.2	6.2	-0.1	-2.6	0.5
2	stable Pleistocene alluvial fans and terraces	-8.6	-6.3	-8.2	-16.7	-23.4	-13.4
3	unstable Pleistocene alluvial fans and terraces	11.6	4.8	17.4	19.3	9.8	4.9
4	recent alluvial fans and riverbeds	-5.3	7.1	0.9	2.8	21.8	11.4
5	*cuvettes*	-9.7	-4.8	-11.1	-13.8	-1.6	-7.2
6	loess formations	5.4	0.3	0.3	3.7	-1.4	1.1
7	resistant rocks	-1.6	-2.4	-5.4	0.5	-2.0	7.3
8	intermediate rocks	-2.1	-2.0	-4.4	-2.2	-3.3	-2.8
9	soft rocks	6.2	0.0	4.4	6.7	2.8	-1.8
	number of sites	**65**	**38**	**85**	**116**	**111**	**95**
	% of total sample	**12.75%**	**7.45%**	**16.67%**	**22.75%**	**21.76%**	**18.63%**

Table 3.24. Weighting of the gain values from table 3.21 by the total number of sites, in order to make a comparison of the absolute importance of the taphonomic map units through time. See also figure 3.4.

3.5. EXTRAPOLATING SITE DENSITIES

Because of the relatively small number of sites per period in the trenched zones (which only occupy 0.4% of the area), and the fact that the area trenched is not fully representative of the total area, no model has been developed for the trenched zones alone. On the other hand, the trenched zones should give the most reliable estimate of site densities possible, because in theory no sites will escape discovery, whereas field walking will only yield those sites showing significant amounts of archaeological remains at the surface. Within the trenched zones we therefore have the most reliable site sample that can be obtained by means of archaeological survey.

Given this reliable sample, it is theoretically possible to perform an extrapolation of the actual amount of sites per map category. The units where most trenches have been dug are the zones of potential sediment accumulation during the Holocene, i.e. the recent alluvial fans, terraces and riverbeds, the colluvial deposits, the unstable Pleistocene alluvial fans and terraces, and the *cuvettes*. In these units, 0.71% of the total surface has been trenched. For these areas, a cautious prediction can be made of the total number of sites to be found. The total number of sites to be expected is simply a multiplication of the number of sites found per area unit and the total area:

$$X = N\frac{x}{n}$$

where

X = the total number of sites;
N = the number of area units;
x = the number of sites found;
n = the number of area units analysed.

However, this extrapolation is not very useful when the error margin is not known. The standard error of the estimate (Shennan, 1988:310) is given by:

$$s_x = N \frac{s}{\sqrt{n}} \sqrt{\frac{1-n}{N}}$$

where

s = standard deviation of the sample.

All probabilistic sampling studies in archaeology depart from the assumption that basic sampling units like survey quadrats (Nance, 1990), or even parcels of land (Kvamme, 1990) can be defined. Casley and Lury (1982:75) state:

'If the total population of the area is very large, compared to the sample to be selected, the variance of, and hence the precision of estimates calculated from the sample data is a function of the absolute number of sample units, not the sampling fraction.'

The basic problem to be solved in order to obtain reliable standard error estimates is therefore defining the size of the sampling units. The smaller these units are, the larger will be the standard errors. In the case of the trenched zones however, the term sampling unit is virtually without meaning, as we can assume that the area trenched has been sampled completely, and therefore no counts per area unit can be performed: the area trenched is equal to one sampling unit. However, with a sample size of 1, the standard errors can not be calculated. The only practical solution – although not an elegant one - is to use the mean surface of the sites as the basic sampling unit. In this case, each individual observation is either a site or a non-site (in statistical terms: we are dealing with a population of ones and zeros), and then the standard deviation of the sample can be calculated with (Shennan, 1988:311):

$$s = \sqrt{\frac{p(1-p)}{n}}$$

where

p = the proportion of interest.

Stratified sampling theory then allows us to narrow down the standard errors for the total sample somewhat by applying the following equation:

$$s_{st} = \sqrt{\sum \left(\frac{N_k}{N} \right)^2 s^2}$$

where

s_{st} = standard error for the complete sample;
N_k = size of stratum k.

The site surfaces involved can of course not be measured with extreme accuracy. Although surface estimates have been made for most of the sites involved, these are given as ranges from 0.0-0.1 ha, 0.1-0.2 ha, 0.2-0.5 ha, 0.5-1.0 ha and so forth. The surfaces given in table 3.25 can therefore only be regarded as rather crude approximations (Epipalaeolithic/Mesolithic not included because of lack of buried sites):

period	**m^2**	**N**
Early Neolithic	3029,4	17
Middle Neolithic	6071,4	28
Late Neolithic	4187,5	40
Bronze Age	1323,5	85
Iron Age	3533,3	45
all periods	***3653,5***	***215***

Table 3.25. Mean site surfaces for all periods except Epipalaeolithic/Mesolithic. N = number of non-visible sites for which a surface estimate was available.

When using the site surface estimates from table 3.25, the standard errors obtained in table 3.26 are relatively large. Obviously, the low number of 'ones' compared to the 'zeros' leads to this large standard error. Because of this, decreasing the number of zeros (large site surfaces; middle Neolithic) is more efficient in reducing the standard error than increasing the number of ones (more sites; Bronze Age).

No.	***Taphonomic unit***		***EN***	***MN***	***LN***	***BA***	***IA***	***ALL***
1	colluvial deposits	**X**	198.7	993.4	0.0	794.7	0.0	**1986.7**
		s_x	*197.2*	*217.9*	*0.0*	*393.7*	*0.0*	***606.8***
3	unstable Pleistocene alluvial fans and terraces	**X**	154.6	154.6	154.6	77.3	77.3	**618.2**
		s_x	*107.6*	*53.7*	*77.8*	*76.2*	*76.1*	***212.7***
4	recent alluvial fans and riverbeds	**X**	1273.0	763.8	1273.0	1527.6	763.8	**5601.2**
		s_x	*562.8*	*218.2*	*407.1*	*618.7*	*436.9*	***1141.0***
5	*cuvettes*	**X**	0.0	0.0	59.9	299.2	59.9	**418.9**
		s_x	*0.0*	*0.0*	*42.5*	*131.2*	*58.7*	***153.8***
	total	**X**	**1626.2**	**1911.7**	**1487.4**	**2698.8**	**900.9**	***8625.0***
		s_{st}	**607.3**	**313.7**	**417.6**	**750.7**	**448.4**	***1321.7***
		%	**37.3**	**16.4**	**28.1**	**27.8**	**49.8**	***15.3***

Table 3.26. Estimated number of sites (X) and standard error (s) for the trenched zones, all periods except Epipalaeolithic/Mesolithic.

The site surface estimations allow us to perform the same extrapolation for the field walked areas. In general, the estimated surfaces for the visible sites are much larger than for the non-visible sites, because of the spread of archaeological artefacts over the surface by erosion and agricultural practices (table 3.27). This obviously means that the number of 'zeros' will be substantially reduced, regardless of the number of sites involved.

The extrapolation of the amount of sites for the field walked zones yields the figures in table 3.28. The total number of sites calculated is much lower than the number obtained for the trenched zones. At the same

time, standard errors are much smaller as well. An extrapolation based on the visible sites in the field walked zones is clearly strongly underestimating the number of sites to be found in the sedimentary zones. Furthermore, an extrapolation based on the larger site surfaces provides 'false security' when it comes to the accuracy of the estimates.

period	m^2	N
Epipalaeolithic/ Mesolithic	78438.6	57
Early Neolithic	73700.0	20
Middle Neolithic	127590.0	50
Late Neolithic	76890.6	64
Bronze Age	38948.7	39
Iron Age	16954.5	11
all periods	***78634.9***	***241***

Table 3.27. Mean site surfaces of visible sites for all periods. N = number of visible sites for which a surface estimate was available[5].

No.	*Taphonomic unit*		*E/M*	*EN*	*MN*	*LN*	*BA*	*IA*	*ALL*
1	colluvial deposits	**X**	201.8	80.7	161.4	161.4	20.2	0.0	**625.6**
		s_x	*56.7*	*37.4*	*49.6*	*51.5*	*19.1*	*0.0*	***83.3***
3	unstable Pleistocene alluvial fans and terraces	**X**	193.5	41.5	193.5	152.0	41.5	13.8	**635.6**
		s_x	*43.2*	*21.8*	*39.9*	*39.3*	*22.0*	*12.8*	***53.5***
4	recent alluvial fans and riverbeds	**X**	36.3	36.3	72.6	36.3	36.3	0.0	**217.7**
		s_x	*35.1*	*35.1*	*49.0*	*35.1*	*35.2*	*0.0*	***83.5***
5	*cuvettes*	**X**	8.5	8.5	0.0	16.9	8.5	8.5	**50.8**
		s_x	*7.4*	*7.4*	*0.0*	*10.5*	*7.5*	*7.5*	***17.7***
	Total	**X**	**440.0**	**166.9**	**427.5**	**366.7**	**106.4**	**22.3**	***1529.8***
		s_{st}	**82.0**	**57.6**	**82.3**	**76.5**	**47.3**	**15.5**	***133.8***
		%	**18.6**	**34.5**	**19.3**	**20.9**	**44.5**	**69.5**	***8.7***

Table 3.28. Estimated number of visible sites (X) and standard error (s) for the trenched zones, all periods.

No.	*Taphonomic unit*	*d (abs)*	*d(%)*	*n(%)*	*d(abs)*	*d(%)*	*n(%)*
1	colluvial deposits	198.67	10.00%	*15.94%*	205.55	10.35%	*15.05%*
3	unstable Pleistocene alluvial fans and terraces	61.82	10.00%	*38.61%*	131.72	21.31%	*12.17%*
4	recent alluvial fans and riverbeds	560.12	10.00%	*6.17%*	419.79	7.49%	*10.47%*
5	*cuvettes*	41.89	10.00%	*48.25%*	105.44	25.17%	*12.83%*
	total area	**862.50**	**10.00%**	***18.60%***	**862.50**	**10.00%**	***12.11%***

Table 3.29. Calculated values of n for an accepted tolerance d of 10%, at the 95.46% confidence limit. Columns 1-3 show the results when d=10% for all map units, columns 4-6 show the results when d is weighed according to map unit size.

[5] The mean estimated site surfaces seem extremely large, especially for the middle Neolithic. The main sites of this period are thought of as central places, and the largest is evaluated to 100 ha. All figures used are confirmed by Jean-François Berger to be realistic estimates of the site sizes observed at the surface.

Theoretically, it is possible to calculate the size of the area that is needed to bring back the standard errors to a more reasonable limit. This is done applying the following equation (Shennan, 1988:310):

$$n = \left(\frac{ZsN}{d} \right)^2$$

where

Z = confidence limit of the estimate in standard deviation units;
d = desired tolerance of the estimate.

Z can be used to define a confidence limit for the results of the equation; a Z value of 2.0 equates to a confidence interval of 95.46%. A finite population correction (n/(1 + n/N)) should be applied afterwards obtain the correct values for the required sample size. The third column of table 3.29 shows the figures obtained when d is set to +/-10% for each single map category. These figures show that it will be necessary to trench about 26 times the area that has currently been trenched in order to obtain a 95.46% reliable estimate within 10% of the total number of sites to be found in the area covered by the four map categories. In general it can be stated that the smaller the number of observed sites in a map unit, the more area needs to be trenched. This means that for statistically reliable estimations of site numbers per map category, the areas with low probabilities of site occurrence should be surveyed more intensively than the areas with high probabilities. In order to reduce the amount of area to be surveyed, while still achieving a tolerance of +/-10% for the whole area, the accepted tolerances for the individual map units may be weighed according to the units' size, as is shown in the last three columns of table 3.29. Of course trenching of such large areas is not feasible – it will therefore be more practical to combine field walking and augering for such an exercise.

3.6. CONCLUSIONS

We set out to investigate the effect of non-random sampling on the interpretation of site quantities and site location distribution in the landscape of the Tricastin-Valdaine region. The results of the predictive modelling have shown that this effect may be very strong indeed. Especially for the later periods, two effects are observed: firstly, the visible sample is not always representative of the sites found in the field walked zones, which means that certain types of sites are easily overlooked during a field walking campaign.

Secondly, the amount of buried sites is very large for the later periods. These sites are found in landscape units that will not yield a comparable amount of visible sites (this is especially true for the recent alluvial plains and riverbeds). An interpretation of site distribution based on the visible sample alone will therefore strongly underestimate the importance of the sedimentary zones for site location.

The actual quantities of sites extrapolated for the sedimentary zones are very large. Basically, it means that the total number of buried sites to be expected in the sedimentary areas is approximately four to seven times as large as the number of visible sites. The actual reliability of the estimate is difficult to judge, given the crude approximation of mean site surfaces that was used to obtain the sampling unit size. However, these mean site surfaces used are not unreasonable estimates, and obtaining more accurate size data will therefore not drastically change the outcome of the extrapolation because of the effect of the small site surfaces when compared to the total area.

The results of both the predictive modelling and the extrapolation strongly emphasize the need for sub-surface surveying methods in sedimentary areas. This need is long recognised in the Netherlands, where augering has become in integral part of archaeological survey in sedimentary areas, and has led to the discovery of many hitherto unknown buried prehistoric sites, sometimes at considerable depth (e.g. Haarhuis, 1995; 1996). The results of the current study indicate that there is no reason to suspect that the situation in France will be very different.

As a last remark, it can be stated that the currently presented predictive model is not very specific for the later periods, as is demonstrated by the lower gain and K_j-values obtained. This means that the model is not very well suited as a tool to guide future surveys, or as an instrument to judge the effect of infrastructural and building activities on the archaeological record. Obviously, the model was not primarily constructed as a tool for archaeological resource management. It served to demonstrate that taphonomy is far more important for site location than was previously thought, and showed the need for a reassessment of both existing site location theory as well as research strategies. A useful model for archaeological resource management is better served by combining elements of the inductive and deductive lines of reasoning, which will probably result in models that make the most of our current archaeological knowledge (Verhagen *et al.*, 2000). This might include further research into the palaeogeography of the area, and the analysis of other site location parameters - which may be different for different archaeological periods or parts of the landscape.

ACKNOWLEDGEMENTS

The authors would like to acknowledge that this work was done in the framework of the ARCHAEOMEDES Project (II), funded by the Directorate General XII of the Commission of the European Union under contract ENV 4-CT95-0159. The authors would like to thank Sander van der Leeuw (Université de Paris I), co-ordinator of the ARCHAEOMEDES Project, for his support. They would also like to thank the following colleagues and institutions involved in the archaeological surveys in the Tricastin and Valdaine regions:

- A. Beeching and J.L. Brochier (CAP Valence), and J. Vital (CNRS) for the assembly and the interpretation of a substantial part of the archaeological database of the Valdaine basin;
- the TGV Méditerranée archaeological project, financed by the French national railway company SNCF (directed by T. Odiot, excavations carried out by AFAN), which contributed more than 60 excavated settlements;
- and the authors of the archaeological map of the region Rhône-Alpes, J.-P. Daugas and C. Laroche (Direction Régionale des Affaires Culturelles de la région Rhône-Alpes).

Finally, the authors would also like to thank Milco Wansleeben (Universiteit van Leiden) for his valuable comments on the predictive modelling methodology.

BIBLIOGRAPHY

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)*. Council of Europe, Strasbourg, pp. 181-190.

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

Beeching, A., J.-F. Berger, J.L. Brochier and J. Vital, 1994. 'De l'espace au territoire en Valdaine, de 9000 à 900 av. J.C.', in: A. Beeching and J.L. Brochier (eds.), *Archéologie spatiale en Vallée du Rhône, espaces parcourus/territoires exploités. Le groupe néolithique et son territoire*. Rapport d'ATP. CAPV, Valence, pp. 38-44.

Berger, J.-F., 1996. *Le cadre paléogéographique des occupations du bassin valdainais (Drôme) à l'Holocène*. Université de Paris I, Paris. PhD-thesis.

Berger, J.-F., J.L. Brochier, C. Jung and T. Odiot, 1997. 'Intégration des données archéologiques et des données naturelles dans le cadre du TGV Méditerranée', in: J.P. Bravard, G. Chouquer and J. Burnouf (eds.), *La dynamique des paysages protohistoriques, antiques, médiévaux et modernes. XVIIe Rencontres Internationales d'Histoire et d'Archéologie d'Antibes.* Éditions APDCA, Antibes, pp. 155-184.

Berger, J.-F., F. Magnin, S. Thiebault and J. Vital, 2000. 'Emprise et déprise culturelle à l'Age du Bronze: l'exemple du bassin valdainais et de la moyenne vallée du Rhône'. *Bulletin de la Société Préhistorique Française* 97:95-119.

Berger, J.-F., 2000. 'Cycles anthropiques et environnementaux à l'Holocène dans des bassins-versants rhodaniens de rang inférieur (Valdaine et Tricastin, Drôme)', in: M. Barrué-Pastor and G. Bertrand (eds.), *Les Temps de l'Environnement, Toulouse: Geode/CNRS, Actes des Journées du Programme Environnement, Vie et Sociétés PIREVS, Toulouse, Octobre 1997*. Presses Universitaires du Mirail, Toulouse, pp. 473-500

Berger, J.-F. and J.-L. Brochier, 2000. 'Evolution des paysages et des climats dans la moyenne vallée du Rhône et sa bordure préalpine de 13000 à 5000 B.P.', in: C. Cupillard and A. Richard (eds.), *Les derniers chasseurs-cueilleurs d'Europe occidentale (13 500 - 5000 av. J.-C.), Actes du Colloque international de Besançon (Doubs, France), 23-25 octobre 1998.* Collection Annales Littéraires de l'Université de Franche-Comté 699. Série Environnement, sociétés et archéologie 1. Presses Universitaires de Franche-Comté, Besançon, pp. 37-58.

Bornand, M., 1967. *Etude pédologique du Bassin Valdainais, 1:25,000*. Institut National de Recherche Agronomique, Service d'Etude des Sols, Centre de Recherche Agronomique du Midi, Montpellier.

Bornand, M., 1971. *Carte des sols de la région de Pierrelatte – Bourg St. Andéol, 1:25,000*. Institut National de Recherche Agronomique, Service d'Etude des Sols, Centre de Recherche Agronomique du Midi, Montpellier.

Brochier, J.L., 1991. 'Environnement et culture : état de la question dans le Sud-est de la France et principes d'étude autour du Chasséen de la Moyenne Vallée du Rhône', in: A. Beeching, D. Binder, J.-C. Blanchet, C. Constantin, J. Dubouloz, R. Martinez, D. Mordant, J.-P. Thévenot and J. Vaquer (eds.), *Colloque international de Nemours, Identité du Chasséen, 1989.* Mémoire Musée de Préhistoire d'Ile de France 4. Paris, pp. 315-326.

Casley, D.J. and D.A. Lury, 1982. *Monitoring and Evaluation of Agriculture and Rural Development Projects*. The Johns Hopkins University Press, Baltimore.

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

Deeben, J., D.P. 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.P. Hallewas (eds.), *Archaeological heritage management in the Netherlands. Fifty years State Service for Archaeological Investigations.* Rijksdienst voor het Oudheidkundig Bodemonderzoek, Amersfoort, pp. 76-118.

Haarhuis, H.F.A., 1995. *De Waalsprong, gemeente Nijmegen. Archeologisch onderzoek fase A1*. RAAP-rapport 122. Stichting RAAP, Amsterdam

Haarhuis, H.F.A., 1996. *Gemeente Nijmegen, De Waalsprong. Archeologisch onderzoek fase A/B deel 2.* RAAP-rapport 175. Stichting RAAP, Amsterdam.

Hodder, I. and C. Orton, 1976. *Spatial Analysis in Archaeology*. Cambridge University Press, Cambridge.

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 – CAA97. Computer Applications and Quantitative Methods in Archaeology 25th Anniversary Conference, University of Birmingham*. British Archaeological Reports, International Series 750. Archaeopress, Oxford, CD-ROM.

Kvamme, K.L., 1989. 'Geographical Information Systems in regional archaeological research and data management', in: M.B. Schiffer (ed.), *Advances in archaeological method and theory I.* University of Arizona Press, Tucson, pp. 139-203.

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. Holos Verlag, Bonn, pp. 257-294.

Leusen, P.M. van, 1996. 'GIS and locational modeling in Dutch archaeology: a review of current approaches', in: H.D.G. Maschner, (ed.). *New Methods, Old Problems: Geographic Information Systems in modern archaeological research.* Occasional Paper No. 23. Center for Archaeological Investigations, Southern Illinois University, Carbondale (IL).

Nance, J.D., 1990. 'Statistical sampling in archaeology', in: A.Voorrips (ed.), *Mathematics and Information Science in Archaeology: A Flexible Framework.* Studies in Modern Archaeology Vol. 3. Holos Verlag, Bonn, pp. 135-163.

Odiot, T., 1985. 'Occupations fortifiées de hauteur en Tricastin', in: A. Duval, (ed.), *Les Alpes à l'Age du Fer. Actes du colloque "Les Alpes à l'Age du Fer", Chambéry.* Revue Archéologique de Narbonnaise, supplément 22. CNRS, Paris, pp. 57-72.

Shennan, S., 1988. *Quantifying archaeology.* Edinburgh University Press, Edinburgh.

Thomas, D.H., 1976. *Figuring anthropology. First principles of probability and statistics.* Holt, Rhinehart and Winston, New York.

Tounsi, I., C. Jung, J.-F. Berger, G. Chouquer, F. Favory and T. Odiot, 1997. 'Etude de la paléohydrographie et du réseau routier ancien en pays tricastin (France, Drôme-Vaucluse) à partir d'images thematic mapper'. *Photo interprétation* 35(1-2):113-126.

Verhagen, P., 1995. 'La carte du potentiel archéologique en Hollande. Une méthode de prédiction fondée sur les données de l'archéologie et du paysage'. *Les Nouvelles de l'Archéologie* 61:34-39.

Verhagen, P., M. Wansleeben and M. van Leusen, 2000. 'Predictive modelling in the Netherlands. The prediction of archaeological values in Cultural Resource Management and academic research', in: Hartl, O. and S. Strohschneider-Laue (eds.), *Workshop 4 Archäologie und Computer 1999.* Forschungsgesellschaft Wiener Stadtarchäologie, Vienna, pp. 66-82. CD-ROM.

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

POSTSCRIPT TO CHAPTER 3

This paper was presented at the Computer Applications and Quantitative Methods in Archaeology (CAA) conference in Ljubljana in April 2000, and appeared in its proceedings in 2001. The research was carried out between 1998 and 2000, and has been repeated in 2002 for the Roman and Medieval period. The study was intended to be published in a French monograph on the TGV-Rhône project, which has failed to appear up to this date. The predictive models have found their way to the DRAC Rhône-Alpes in Valence.

It was Jean-François Berger's idea to use predictive modelling to analyze the differences between the pre-1990s theories on site location in the Tricastin-Valdaine area, and the new insights that appeared as a result of the Archaeomedes en TGV-Rhône projects. His research made clear that especially the fluvial and colluvial deposits in the area had been systematically ignored in the archaeological record. Berger *et al.* (1997) estimated that between 30 and 50% of the archaeological sites in piedmont and valley bottom zones in the Rhône Valley is located deeper than 2 to 5 m below the surface (see also Raynaud, 2000). Jean-François Berger's enormous geo-archaeological experience in the area made it possible to reclassify the geological and pedological maps of the area into what he called 'taphonomic units'. This resulted in a classification of the area in stable, and unstable, dynamic environments. The stability of the landscape was thought to be the principal determining factor for explaining site location in the area. As the predictive modelling demonstrated, this is certainly true for the earlier prehistoric periods. However, for the later periods weaker correlations were found between site location and the various taphonomic map units. Nevertheless, the initial suspicion that sedimentary zones were underrepresented in the 'traditional' data set was fully confirmed by the modelling. Even so, we have to keep in mind that the visible site sample may not be a representative sample either, because of differential artefact visibility. It is for example well known that repeated field walking in the same area will result in the discovery of sites that were not noticed in previous surveys, because of the effects of changing weather conditions and the different stages of crop cultivation.

From a methodological point of view, the paper tried to compare different quantitative techniques used for predictive modelling. Looking back, it was insufficiently clear to me that, from the techniques chosen, three are methods of model performance optimization, whereas the χ^2-test is a statistical method of testing if

the site distribution found differs from a 'by chance'-model (see chapter 7 for further discussion on statistical testing methods in predictive modelling). It can only be used to decide whether a variable should be included in the model at all. In order to have achieved statistically more sophisticated models, Monte Carlo simulations should have been used (see also chapter 7 on the issue of resampling), but at the time no 'off the shelf' GIS software was available to me that could have done that. Furthermore, the use of the 99.9% confidence level with the χ^2-test in the paper is too strict, and the threshold for rejecting the hypothesis of a 'by chance'-distribution therefore placed too high. A 95% confidence level suffices for normal statistical practice, and should be enough for predictive modelling purposes as well. For certain periods, the conclusion that the distribution is not significantly different from a 'by chance'-distribution should therefore be taken with the proverbial grain of salt.

Despite these shortcomings, the modelling made clear that the use of gain development graphs is an easy way to perform a classification into zones of low, medium and high potential, and demonstrated the point that biased archaeological data sets will lead to biased predictive models. Perhaps the most interesting aspect of the modelling however, is the attempt made in the last paragraph to calculate the number of sites that might still be hidden beneath the soil. Even though the base data were far from ideal because of the absence of reliable site surface estimates, the staggering totals obtained by using very straightforward and standard statistical procedures constitute a clear warning to users of predictive models anywhere that empty regions simply do not exist.

ADDITIONAL REFERENCE

Raynaud, C., 2000. 'Territoire et Peuplement en France, de l'Age du Fer au Moyen Age. L'Archéologie Spatiale à la Croisée des Chemins', in: J. Bintliff, M. Kuna and N. Venclová (eds.), *The future of surface artefact survey in Europe.* Sheffield Archaeological Monographs 13. Sheffield Academic Press, Sheffield, pp. 57-71.

CHAPTER 4 Quantifying the Qualified: the Use of Multicriteria Methods and Bayesian Statistics for the Development of Archaeological Predictive Models[1]

4.1. INTRODUCTION

Over the past ten years, archaeological predictive modeling in the Netherlands has been the subject of a sometimes-heated debate (see Verhagen *et al.*, 2000). After seminal publications by Wansleeben (1988), Ankum and Groenewoudt (1990), and Soonius and Ankum (1990), a number of predictive maps have been produced by public archaeological institutions in the Netherlands (RAAP[2] and ROB[3]). At the same time, academic archaeologists have studied the methodological and theoretical aspects of predictive modeling, and criticized the modeling concepts used in public archaeology in several publications (Wansleeben and Verhart, 1992; van Leusen, 1993; 1995; 1996; Kamermans and Rensink, 1999).

Figure 4.1. Location of the study area in the Netherlands.

The inferential or inductive approach, already criticized in Brandt *et al.* (1992) for its inability to cope with the low quality of many archaeological datasets, has gradually been replaced by a more intuitive way of model development, trying to make the best of both worlds by including quantitative data when they are available. Coupled to this development towards more deductive mapping, it is notable that the multi-variate approach has been replaced by the use of a reduced number of variables, that are supposed to have the strongest predictive power for a particular region or archaeological period. These models can best be characterized as 'hybrids', and essentially are descriptions of existing knowledge, rather than extrapolations. A consequence of this approach is that error margins and uncertainties are never specified, and on the whole the models lack a clear formalized methodology for including both 'hard' and 'soft' knowledge.

The current chapter discusses the possibilities of improving the applied methodology by focusing on the formalization of the inclusion of 'expert judgment' or subjective knowledge into the mapping. Two

---

[1] This paper also appeared in Mehrer, M. and K. Wescott (eds.), 2006: *GIS and Archaeological Site Location Modeling*. CRC Press, Boca Raton. pp. 191-216.

[2] RAAP is a private archaeological consultancy firm, specialized in archaeological survey and predictive modeling

[3] Rijksdienst voor het Oudheidkundig Bodemonderzoek, the Dutch national archaeological service

methodological innovations are suggested for this: the application of a multicriteria decision-making framework to the modeling, and the use of Bayesian statistics for developing the knowledge base needed for the model. This methodological framework is tested on a case study in the municipality of Ede (figure 4.1), where the model of Soonius and Ankum (1990) was originally applied.

4.2. MULTICRITERIA DECISION MAKING AND ITS RELEVANCE TO PREDICTIVE MODELING

Multicriteria decision making (MDCM) is a set of systematic procedures for analyzing complex decision problems. By dividing the decision problem into small, understandable parts, and then analyzing these parts and integrating them in a logical manner, a meaningful solution to the problem can be achieved. Decision making includes any choice among alternative courses of action and is therefore of importance in many fields in both the natural and social sciences. These types of decisions usually involve a large set of feasible alternatives and multiple, often conflicting and incommensurate evaluation criteria. Archaeological predictive modeling fits into this framework, as it is a way to evaluate the archaeological potential of an area, and provide the basis for decision making in prospection design as well as in planning procedures. In fact, it is recognized as such by Kvamme (1990): 'A [predictive] model is a *decision rule* conditional on other non-archaeological features of locations' [emphasis added].

Many archaeologists may not be familiar with the concepts and terminology used in MCDM, so a condensed description of its core notions follows here. This description closely follows the outline presented by Malczewski (1999); a slightly different terminology can be found in other publications, like Nijkamp *et al.* (1990).

MCDM can be broken down into the following components (Malczewski, 1999:82):

- The definition of a **goal** the decision maker attempts to achieve;
- The selection of a set of **evaluation criteria** (called objectives or attributes);
- The **decision maker** and his or her **preferences** with respect to the evaluation criteria;
- The definition of a set of **decision alternatives**;
- The calculation of a set of **outcomes** associated with each attribute / alternative pair.

DEFINING GOALS

A goal is a desired state of affairs. In a predictive modeling context, the goal can for example be defined as minimizing the impact of planning measures on the archaeological record. A similar goal could be a maximum reduction of the costs associated with archaeological investigations in the area under consideration. The defined goal may be broken down into several *objectives*, which can best be thought of as intermediate goals. This conceptual framework forms the basis of the Analytical Hierarchy Process (Saaty, 1980), a widely used method for MCDM. By defining objectives, a hierarchical structure of decision making can be developed. Decisions can then be made by comparing objectives, or at the lowest level of the hierarchy, by comparing *attributes*. Attributes are measurable quantities or qualities of a geographical entity, or of a relationship between geographical entities. These are the basic information sources available to the decision maker for formulating and achieving the objectives. An attribute is a concrete descriptive variable; an objective is a more abstract variable with a specification of the relative desirability of that variable. The attributes used for

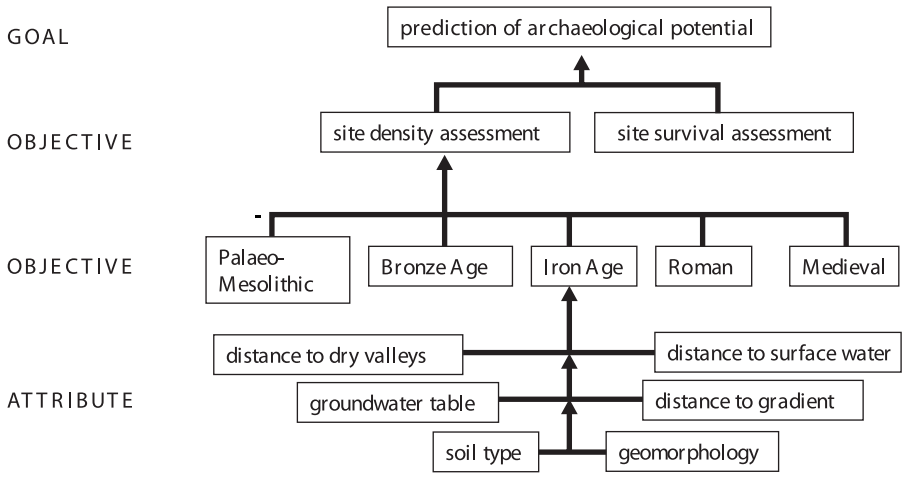


Figure 4.2. The Ede predictive model (Soonius and Ankum, 1990) and its hierarchical structure of objectives.

archaeological predictive modeling are usually a limited number of environmental variables; they can be supplemented with expert knowledge on less measurable variables.

Given the goal of minimizing the impact of planning measures on archaeology, one of the objectives under consideration might be the selection of areas of minimal site density, and another the selection of building methods that are not damaging to the archaeological remains. In this example, predictive modeling is only a way of defining an objective, rather than a way of achieving an immediate goal. Archaeological predictive modeling is a typical example of multiattribute decision making (MADM), that obtains preferences in the form of functions and weights directly for the attributes. MADM problems are those that have a predetermined, limited number of alternatives, as is common in environmental impact assessments where, for example, a limited number of railway alignments may have to be compared. However, in a GIS-context, each single raster cell or polygon can be seen as a decision alternative, as outcomes are calculated for each entity.

When applying the MCDM framework to a predictive modeling study, such as was done in Ede by Soonius and Ankum (1990), it is easy to break down the model into a hierarchical structure of objectives (figure 4.2). The final model presented is aimed at predicting the potential of every single raster cell for finding undisturbed remains of prehistoric settlements. It has two objectives at the highest level: an assessment of the chance of survival of archaeological remains, and an assessment of overall site density. The assessment of overall site density is achieved by combining five intermediate objectives: determining site density for five separate archaeological periods. Each of these subobjectives is in turn evaluated using six attributes, selected by means of a χ^2 test.

SELECTION OF THE EVALUATION CRITERIA

In this phase of MCDM, the attributes to be used are specified, and a measurement scale is established. This is equivalent to the selection of the predictor variables and establishing their values, for example, in terms of site density or probability. Attributes may be used for different objectives, possibly with other values attached.

Attributes should be *complete, operational, decomposable, nonredundant,* and *minimal*. A set of attributes is complete if it covers all relevant aspects of the decision problem and indicates the degree to which the overall objective is achieved. One of the enduring criticisms on archaeological predictive modeling is that the set of attributes used is not complete; especially social and cultural variables are assumed to be missing from the full set of attributes (see also chapter 10). A set of attributes is operational if it can be used meaningfully in the analysis (it is understandable), so decision makers can understand the consequences associated with alternative decisions. Furthermore, an attribute should be *comprehensive* (have a direct relation to the decision problem) and *measurable*. This last condition in many cases implies choosing a proxy attribute; so instead of 'a dry piece of land to live on', the more tractable variable 'groundwater table' might be employed. The use of proxy attributes implies that there is a missing link between the information available and the information necessary (Beinat, 1995).

In the model created by Soonius and Ankum (1990), the environmental variables used were selected by means of a χ^2 test using the known archaeological sites. As has been shown in van Leusen (1996), the actual application of the test was not done correctly. The violations of the statistical assumptions for the Ede model are, however, not too grave as long as the analysis is not considered per period. Even with Yates's correction applied the values of χ^2 are high enough to justify the selection of the analyzed variables for the predictive mapping on methodological grounds. However, this does not mean that the set of attributes analyzed is complete, although quite a number of environmental descriptors were analyzed.

Furthermore, it may be suspected that some of the variables used are redundant, as no independence check was performed. Spatial cross-correlation indices (Goodchild, 1986) can be calculated for pairs of raster maps and used as a first measure of the independence of the attributes used. The calculation should be based on the attribute values associated with the different maps. Correlation ranges from +1 to -1. A positive correlation indicates a direct relationship between two layers, such as when the cell values of one layer increase, the cell values of another layer are also likely to increase. A negative correlation means that one variable changes inversely to the other. A correlation of 0 means that two layers are independent to each other.

$$c = \Sigma\, c_{ij} / (\sqrt{(\Sigma\, (z_i - \bar{z}_i)^2)} * \sqrt{(\Sigma\, (z_j - \bar{z}_j)^2)})$$

where

$c_{ij} = (z_i - \bar{z}_i) * (z_j - \bar{z}_j)$ or the similarity of i and j's attributes;
z = attribute value;
i = any cell in grid 1;
j = any cell in grid 2 (on the same location).

From table 4.1 it is clear that at least two variables should be regarded with suspicion for use in the predictive model, i.e. groundwater table and distance to dry valleys, both of which seem to be somewhat correlated to the geomorphological map. This is not very surprising, as the groundwater table is related to

elevation, and the distance to dry valleys was obtained by extracting the dry valleys from the geomorphological map. A third possible correlation can be observed between the soil map and the distance to ecological gradient. This seems to indicate that certain soil types are related to the presence or absence of an ecological gradient.

	1	2	3	4	5	6
1	**1.00000**	0.17658	0.10074	0.18295	0.33574	-0.11646
2	0.17658	**1.00000**	0.55118p	0.43743	-0.01326	-0.08668
3	0.10074	0.55118	**1.00000**	0.36713	-0.15656	-0.09246
4	0.18295	0.43743	0.36713	**1.00000**	0.16368	-0.14212
5	0.33574	-0.01326	-0.15656	0.16368	**1.00000**	-0.07790
6	-0.11646	-0.08668	-0.09246	-0.14212	-0.07790	**1.00000**

Table 4.1. Correlation matrix for all six variables use by Ankum and Soonius (1990). 1 = soil type; 2 = geomorphological unit; 3 = groundwater table; 4 = distance to dry valley; 5 = distance to ecological gradient; 6 = distance to surface water.

DEFINING MEASUREMENT SCALES

Measurement scales can be obtained by using normalization, value (or utility) functions (Keeney and Raiffa, 1976); probabilistic methods; and fuzzy set membership (e.g., see Burrough, 1989 and Burrough *et al.*, 1992). A distinction can be made between objective probability (like site-density measures) and subjective probability. The latter is also known as 'prior belief' in the context of Bayesian statistics.

The objective probability approach received a lot of attention in archaeological predictive modeling in the 1980s, and in particular the use of multivariate statistical techniques like logistic regression that can be used to obtain objective weights (e.g., see Warren, 1990). However, in many cases the weights obtained by multivariate statistics are not as objective as one would like them to be. As was noted at the outset, the use of quantitative, multivariate methods for predictive modeling in the Netherlands has been replaced over the past 10 years by qualitative or semiquantative methods. The main reason for this is the fact that frequency statistics have not been able to deliver their promise of objective predictions because of the poor quality of the archaeological data sets used. Many archaeological data sets are biased towards certain site types that have been recorded under specific terrain conditions, so there will be many situations where the available archaeological data set can not be used as a representative sample of the target population. In those cases, expert judgment or subjective belief may be used to estimate map weights.

DEFINING PREFERENCES

The decision maker's preferences with respect to the evaluation criteria should be incorporated into the decision model; they express the importance of each criterion relative to other criteria. The decision maker is simply the person(s) involved in trying to achieve the defined goal. Ideally, experts provide facts and decision makers values (Beinat, 1995). In predictive modeling, however, the decision makers are often also the experts. They might judge that, for the prediction of Bronze Age graves, other evaluation criteria should be used than for Medieval settlement locations. Similarly, different preferences can be specified for the Bronze

Age graves and Medieval settlements when making a decision on what to investigate under the constraints of the available budget.

A number of methods are available to obtain a numerical representation of preferences. Of these methods, pairwise comparison (Saaty, 1980) seems to be the most effective, but ranking methods and rating methods are easier to apply. The important element is that these methods are all meant for the comparison of value judgments, and as such involve expert opinion and subjective reasoning.

The definition of the preferences, in fact, takes place at *all* hierarchical levels of the decision-making process: attributes can be compared, but objectives can be as well. For example, the decision maker might decide that groundwater table is more important for site location than soil texture. After finishing this evaluation, he or she might decide that Neolithic sites are not as important as Roman sites. And after this evaluation, he or she might decide that site density is a less important criterion than site preservation. In this way, a nested hierarchy of decision making is created, in which all decisions can be subjected to the same cycle of criterion selection, establishment of measurement scales, definition of alternatives and preference definition (the Analytic Hierarchy Process). It should be noted that the definition of preferences can be avoided only when the criteria used are truly independent, and can be measured in terms of objective probability (for example in a logistic regression equation when all statistical requirements have been fulfilled). In all other cases, value judgments will be necessary to weigh the evaluation criteria.

The procedures of establishing measurement scales and criterion preferences, as for example applied by Dalla Bona (1994; 2000), who refers to it as the 'weighted value method', and even of nesting objectives are of course not new in archaeological predictive modeling (e.g., Kohler and Parker, 1986, who distinguish a Hierarchical Decision Criteria Model). However, they are usually not recognized as belonging to the more generic class of MCDM methods. Van Leusen (1993) for example suggested that 'translating' archaeological intuition into weighting schemes and other types of classification rules and embodying them in an expert system would at least make 'intuitive' approaches reproducible, which seems an adequate description of applying MCDM methods. It should be noted, however, that this methodology, which heavily relies on expert judgment, has not been the dominant one in (especially American) literature on the subject up to today, as is illustrated by most of the applications found in Wescott and Brandon (2000). An example of the rather suspicious attitude towards subjective weighting can be found in the paper by Brandt *et al.* (1992), who rejected the use of purely deductively derived weights for map layers and features, basically because the current state of archaeological theory would not be able to give more than rough notions about human locational behavior. However, at the same time, these authors could not achieve complete reliance on inductive methods either (van Leusen, 1996), leading to the 'hybrid approach' that was also used by Ankum and Soonius (1990). Recently, some interest can be observed in the use of land-evaluation methods as a purely deductive technique for predictive modeling (Kamermans, 2000), an approach which has been applied previously in archaeological studies of prehistoric land use (Kamermans, 1993; Finke *et al.*, 1994). However, one should not necessarily equate deductive modeling with expert judgment weighting: the experts usually arrive at their judgment through a combination of deductive and inductive arguments.

ESTABLISHING THE DECISION RULES

This phase brings together the preceding three steps for the overall assessment of the alternatives (ranking of alternatives). The most commonly applied method is simple additive weighting, also known as

weighted linear combination. Similarly, probabilistic additive weighting can be used to obtain a ranking of alternatives. The prerequisite of all addition methods is that the attributes used be conditionally independent.

Kohler and Parker (1986) review four different types of decision rules that can be applied to archaeological predictive models. The Fatal-Flaw Decision Criteria Model is the most constraining of these; it results in a binary response (yes or no). In MCDM, this decision rule is known as a noncompensatory method called *conjunctive screening*. Under conjunctive screening an alternative is accepted if it meets specified standards or thresholds for *all* evaluation criteria (Boolean AND). The Hierarchical Decision Criteria Model is a variant of this; it performs a conjunctive screening as well, but each time for a different objective. Conjunctive screening is also found in the application of land evaluation, where a land unit can only be classified as being suitable for certain kinds of cultivation if it meets all evaluation criteria. *Disjunctive screening* on the other hand accepts the alternative if it scores sufficiently high on *at least one* of the criteria (Boolean OR). This is a method not usually found in archaeological predictive modeling.

If no direct binary response is desired, compensatory methods are applied. They require a value judgment for the combination of (possibly conflicting) criteria, and can be applied only when all the evaluation criteria are measured in the same units. Afterwards, a constraint can be placed on the outcome of the decision rule, for example to distinguish 'crisp'[4] zones of high and low probability. Kohler and Parker (1986) distinguish the Unweighted Decision Criteria Model and the Weighted Additive Decision Criteria Model. The unweighted model is in fact a special case of additive weighting, as all weights are equal - which is a value judgment in itself.

The combination of nonindependent attributes can by done by means of multiplication. This is to be avoided in most cases, as it implies that the interactions between the attributes are known, and these must then serve as the input for the multiplication equation.

4.3. BAYESIAN STATISTICS AND PREDICTIVE MAPPING

COMBINING OBJECTIVE AND SUBJECTIVE WEIGHTS

It is important to note that between the extremes of purely objective and purely subjective weighting, compromises of the two can be used. Two routes can then be followed: the first one is by starting with an objective weighting, and the weights are adapted afterwards by consulting experts. This is basically the method employed by Deeben *et al.* (1997). This form of combination lacks a set of formal rules for application, and will therefore not produce a transparent model unless all modifications to the objective weighting are clearly specified. The second route, which is further explored in this chapter, starts with a subjective weighting, and uses any quantitative data available to modify the weights, but only when the data are considered to be a representative sample. In essence, this is the concept of Bayesian statistics, where a subjective prior belief is modified using quantitative data to obtain a posterior belief:

*posterior belief = conditional belief * prior belief*

[4] sharply defined

The few published applications of Bayesian statistics in predictive modeling inside (van Dalen, 1999) and outside archaeology (Aspinall, 1992; Bonham-Carter, 1994) have one thing in common: the assumption of a uniform prior probability for all map categories. This assumption is the simplest possible form of formulating prior beliefs, and equates to a situation where no prior information is available. In the studies mentioned, most attention is paid to the establishment of the conditional beliefs. It can be shown that the p_s/p_a[5]–ratio (a commonplace indicator of site density that is also used by Deeben *et al.* (1997)) is equivalent in a Bayesian context to the ratio of prior to conditional probabilities under the assumption of a uniform prior probability (see Buck *et al.*, 1996). This is the reason that Bonham-Carter (1994) uses p_s/p_a ratios as well for the development of a Bayesian geological predictive model. Similarly Aspinall (1992), in an ecological application of Bayesian statistics, comments that conditional probabilities can be expressed as relative frequencies of occurrence. If the condition of independence of the variables is met, the calculation of posterior probabilities is then simply a question of multiplying the p_s/p_a ratios per variable with the prior probabilities. When using a logarithmic normalization the posterior probabilities can even be calculated by simple addition instead of multiplication (Bonham-Carter, 1994).

The uniform prior probabilities are based on the currently found site density per area unit. When combined with p_s/p_a ratios used as conditional probabilities, the sum of the predicted posterior probabilities per area unit should equal the number of observed sites, as p_s/p_a is a dimensionless number indicating relative site densities. Bonham-Carter (1994) notes that the ratio of observed to predicted sites can therefore serve as a measure of violation of the assumption of conditional independence of the variables. However, it may be desirable not to predict an absolute number of sites, as the size of the target population is not known. In that case, the prior probabilities should be normalized on a scale of 0 to 1. When using a uniform prior probability, this implies that any Bayesian predictive modeling exercise equates to calculating the p_s/p_a ratios per map category.

Apart from the uniform probability distribution, Buck *et al.* (1996) show a number of other distributions, either discrete or continuous, that can be used to model both prior (Orton, 2000) and posterior probabilities. In the case of categorical maps, a binomial distribution can be implemented by breaking down the nominal variables into binary ones (Bonham-Carter, 1994). The prior and posterior belief for each map category can subsequently be modelled by means of a (continuous) Beta distribution (Buck *et al.*, 1996), allowing for the establishment of standard deviations around the mean of the posterior belief. The form of the Beta distribution is directly dependent on the sample size and the conditional probability derived from it. The larger the sample size, the smaller the standard deviations will become, and the closer the mean of the distribution will be to the conditional probability found. A useful aspect of Beta distributions is the fact that they will also yield a mean and a standard deviation in cases where no sites have been found on the map unit of interest. They can be used when the available sample is small.

FORMULATING THE PRIORS

The following question to be answered is: if we do not want to use an assumption of uniform probability, how can we establish prior probabilities based on expert judgment? This brings us back into the realm of multicriteria decision making, and, precisely, the issue of specifying preferences by the decision maker. Three basic methods can easily be applied:

[5] p_s = proportion of sites found in map unit X; p_a = proportion of area taken up by map unit X

Ranking methods. The decision maker expresses a ranking of the criteria under consideration, and the following equations may then be used to obtain numerical weights from the rank-order information (rank sum weights):

$$w_j = n - r_j + 1 / \Sigma (n - r_k + 1)$$

or (rank reciprocal weights)

$$w_j = (1 / r_j) / \Sigma (1/r_k)$$

where

w = weight for criterion j
n = number of criteria under consideration
r = rank position of the criterion

In general, the larger the number of criteria to be ranked, the less useful the method will become.

Rating methods. The decision maker tries to estimate weights on a 1 to 100 scale (or any other conceivable numerical scale). This is the most widely applied method in archaeological predictive mapping (e.g., Dalla Bona, 1994; 2000).

Pairwise comparison method (as part of the Analytical Hierarchy Process; Saaty, 1980). Criteria are compared in pairs, and intensities of importance are attributed to each pair. It can only be performed if all criteria are measured on the same scale. It is suggested by Malczewski (1999) that the method is the most effective technique for spatial decision making, and as such should receive more attention. A good introduction to the technique is given in Eastman *et al.* (1993). The number of comparisons involved can become very large if the number of criteria increases. Comparisons are made in linguistic terms (table 4.2). A maximum number of nine comparison levels is given that can be related to an intensity number. An intensity number of 1 implies equal importance (the diagonal in the comparison matrix), an intensity of 9 is translated as extreme importance. This means that when a zone of high archaeological value is judged to be ‘extremely more important’ than a zone of low archaeological value, that a value of 9 is given to the ‘high compared to low’ cell in the comparison matrix. A value of 1/9 should then be given to the ‘low compared to high’ cell, resulting in a reciprocal matrix.

In essence, the procedures described above solve the problem of formalizing the use of subjective information in a predictive map.

intensity of importance	*definition*
1	*equal importance*
2	*equal to moderate importance*
3	*moderate importance*
4	*moderate to strong importance*
5	*strong importance*
6	*strong to very strong importance*
7	*very strong importance*
8	*very to extremely strong importance*
9	*extreme importance*

Table 4.2. The scale used for pairwise comparison. Source: Malczewski (1999).

BAYESIAN STATISTICS AND INDUCTIVE LEARNING

The available methods for preference specification demand that the expert express his or her preferences in a numerical manner, or at least be able to provide a preference ranking. One might therefore ask: is it of any use to specify preferences in a numerical way, apart from specific cases where a numerical answer is desired? Two arguments can be given in favor of using numerical preferences:

The measurement scales should be comparable when applying decision rules; some form of quantification is necessary to judge the outcome of any multi-criteria decision making process.

The fact that predictive maps will be nothing but a representation of current knowledge makes it necessary that the models produced be amenable to improvement, i.e., the models should be able to learn. To achieve this, the modeling procedure should be transparent and reproducible; this is more easily done using numerical preferences than by using linguistic terms.

The second point may need some clarification. Let us take the case where a map unit has been qualified as having a low archaeological value. The definition of 'low value' in this case may imply a relatively low density of archaeological finds. However, the linguistic definition does not include an assessment of the actual quantities or probabilities involved. Suppose that, in the course of several years, a number of new settlement sites are found within this particular unit, e.g., because of the use of new prospection methods. When do we decide that the archaeological value of this particular unit no longer is low but should be intermediate or high? Without a numerical decision rule, the interpretation of the archaeological value of the unit is neither transparent nor reproducible.

Advocates of Bayesian statistics are always very confident that it offers a way of 'inductive learning'. Venneker (1996) for example states that "it constitutes a computationally efficient recursive process in which the entire data stream is captured in the posterior belief of a hypothesis and need not be recalculated each time new or additional independent data become available". The bottleneck in this statement is the fact that the new data should be independent from the old data in order to be able to adapt the posterior belief; otherwise there is the very real danger of self-fulfilling prophecies. This is precisely the problem of using archaeological predictive maps for guiding surveys: there will be a natural tendency to select those areas where high site densities are predicted, and this may lead to an ever-increasing amount of biased-sample data. Even though the Bayesian approach has the appeal of being a formal method applicable in a rather straightforward way to data that are far from optimal, the approach itself does not solve the problem of using biased data. One advantage of

Bayesian statistics, however, is that if the new data are collected independently, there is no need for random sampling (Orton, 2000); representative sampling is good enough – a task that may be difficult enough in itself.

4.4. APPLICATION: THE PREDICTIVE MAP OF EDE

In order to illustrate the approach outlined above, a case study was performed using the new predictive map of the municipality of Ede, that was recently made by RAAP (Heunks, 2001; figure 4.3). This map was commissioned by the municipality to replace the 1990 map, which had become outdated. The new map is primarily based on a qualitative interpretation of the 1:50.000 soil map of the area. Map units were combined in order to arrive at what can best be described as 'archaeological land units' that were subsequently evaluated for their archaeological value (table 4.3). In essence, we are dealing with a single-attribute map that can be evaluated for different criteria, like site density per period or site-preservation conditions.

The prior probabilities were obtained by contacting four experts on the archaeology of the region. Of these, one refused to cooperate on the grounds that the land-units map alone could not be used for a predictive model as it does not try to take into account social and cultural factors that might have influenced site location. The other three experts were willing to evaluate the map units using the ranking, rating and pairwise comparison methods that were outlined above for both site-density and site preservation potential.

The responses received indicated that the rank reciprocal method will result in inconsistent weights when compared to the other three methods. There is no evidence that any of the remaining three methods used performs better than the other ones. Pairwise comparison is without any doubt a lot more time consuming, even though it is theoretically the most appropriate and efficient (Malczewski, 1999). The main reason for applying pairwise comparison is to obtain weights that are independent of the overall weighting that is obtained with the rating and ranking method. However, when confronted with all three methods, the experts took the exercise as a test on maintaining consistency between methods, resulting in a relatively consistent weighting between methods. Apart from that, the experts were hesitant to indicate differences between units they were uncertain about, preferring to attribute equal weights instead (or in Bayesian terms, specifying uninformative priors).

Comparison of weights between experts showed that some disagreement exists on the importance of the land units, both for site-density as well as for site-preservation potential. Whereas one expert took soil type as the most important factor influencing site density, a second one evidently believed that geomorphology was more important. There was also a strong disagreement on the importance of the partly deflated drift-sand areas for site preservation. This may have been the consequence of not fully explaining the map legend to the experts. This particular unit was thought to be important in terms of site preservation, as drift sand may have covered the previously existing surface. In fact, site-preservation potential is influenced by two factors - the presence of a soil cover and the groundwater table - and one expert clearly believed that a high groundwater table was much more important, possibly reflecting a preference for well-conserved organic remains. Of course this brings us to the question of who to believe. As it will in most cases be impossible to weigh the experts' r on a scale of reliability of response, the obvious solution is taking the mean weight of the experts' responses as the prior belief (Beinat, 1995). This also implies that is possible to calculate the variance of the experts' responses. An alternative option is to consider each expert's opinion as an independent sample.

number	land unit
LOW LYING SANDY PLAIN (GELDERSCHE VALLEI)	
1	low ridges and hillocks
2	low ridges and hillocks with plaggen soils
3	undulating plains
4	valleys and depressions
LATERAL MORAINIC HILLS (VELUWE)	
5	Late Pleistocene aeolian sands with moder podzol soils
6	Late Pleistocene aeolian sands with plaggen soils
7	Late Pleistocene aeolian sands with humic podzol soils
8	fluvial and periglacial sands with moder podzol soils
9	fluvial and periglacial sands with plaggen soils
10	fluvial and periglacial sands with humic podzol soils
11	depressions
12	stream valley
13	drift sands, both covered and deflated areas
14	drift sand, deflated areas

Table 4.3. Archaeological land units, used as the basis for the Ede predictive map (Heunks, 2001).

SITE DENSITY

For the site density mapping, a Beta-distribution was used to model the prior and posterior belief. The Beta-distribution has the following form (for a more detailed description, see Buck *et al.*, 1996):

$$C * p^{a-1} * (1 - p)^{b-1}$$

where

p = the proportion of sites in unit X;
1 – p = the proportion of sites not in unit X;
a – 1 = the number of 'successful draws' from a sample of size n;
b – 1 = the number of 'unsuccessful draws' from a sample of size n (b = n + 2 - a); and
C = a normalizing constant dependent on a and b.

The prior form of the Beta distribution is obtained by setting *a* to 1, as no sample data are available (Orton, 2000). The value of *b* and the corresponding standard deviation can then be calculated directly, as the mean of the distribution (the weight attributed by the expert to the map category) is given by $a / (a + b)$, and the variance by $ab / [(a + b)^2 (a + b + 1)]$ (Buck *et al.*, 1996). Once sample data become available, the Beta-distribution can be updated by simply adding the number of 'successful' and 'unsuccessful' draws to $a - 1$ and $b - 1$, respectively. Eventually, the mean of the distribution will move closer towards the mean of the actual sample, and standard deviations will decrease with increasing sample size. In this particular case we are dealing with three experts. If each of these experts' opinions is treated as an independent sample, the total value of *a* increases to 3, and *b* to the sum of *b* for each independent sample.

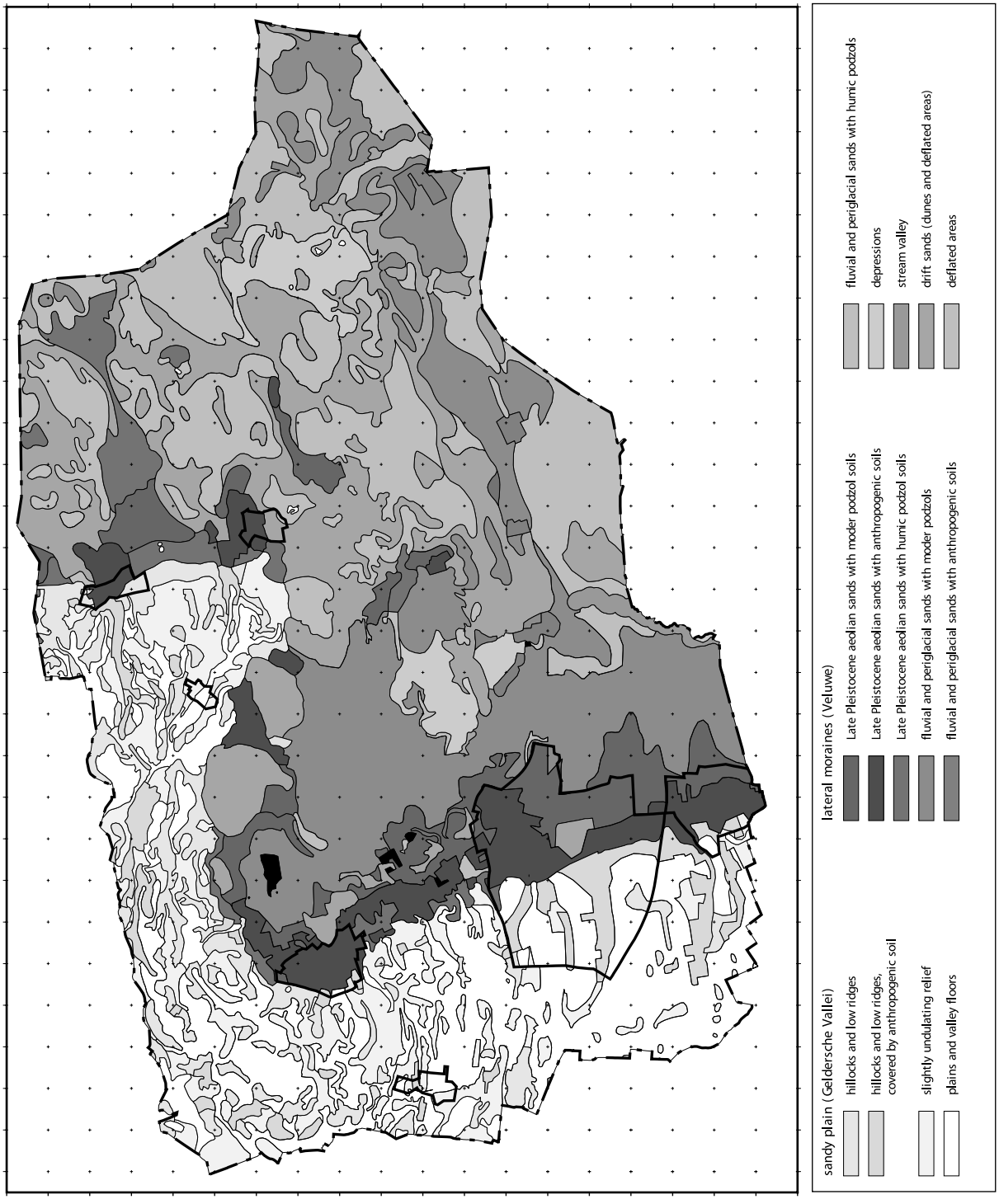


Figure 4.3. The land units maps of the municipality of Ede (Heunks, 2001).

Alternatively, by taking the mean of all experts' opinions, and calculating the variance of the responses, one can obtain values for *a* and *b* using the equations given in Robertson (1999)[6]. This will result in a radically different outcome for the prior-probability distribution, and consequently will have a profound effect on the influence of the conditional probabilities on the posterior belief. This method will attribute much more weight to the experts' opinions (especially when they are in agreement), whereas the first method will tend to emphasize the importance of the incoming new site data.

When setting *a* to 1 for formulating the prior probabilities, the relatively large number of 235 known sites in the municipality of Ede leads to posteriors that are very close to the conditional probabilities. The experts' judgment will become quickly less important once considerable amounts of data become available. Unfortunately, it seems improbable that the data set can be considered a representative sample of the total area. The most important reason for this seems to be a research bias of the registered archaeological data set towards areas that are open to field survey. However, large areas of the municipality are covered in woodland or grassland, and even though no quantitative data are currently available to estimate the importance of this effect, it can be suspected that certain land-use types are related to specific land units.

The use mean and standard deviations of the experts' opinions to define prior probabilities leads to posteriors that are closer to the prior probabilities in cases where the experts more or less agree. Where strong disagreement exists (larger standard deviations), the posterior probabilities are closer to the conditional probabilities. Whether this alternative method of prior formulation is preferable is probably dependent on the importance attributed to the available site data. Given the fact that the existing site sample is viewed with suspicion, in this particular case it may be the best method for formulating priors.

It is possible to determine confidence intervals around the prior and posterior mean, but as we are not dealing with a normal distribution, the standard deviations cannot be used to create confidence intervals; these must be calculated in a statistical package. Plotting the prior distribution together with the posterior distribution shows how far the posterior beliefs are removed from the prior beliefs. Tables 4.4a and 4.4b show how the model develops if all available site data in the municipality of Ede are included to obtain posterior beliefs for both methods of prior formulation.

SITE PRESERVATION POTENTIAL

The initial goal of finding the areas with a minimal number of well-preserved sites was broken down into two objectives: mapping of site-density and site-preservation potential. For both criteria the same attribute (the land-units map) was used, but different weights were applied. However, when the two objectives are combined it is clear that site-density and site-preservation potential are not independent objectives that can simply be added to arrive at a final weight. Indeed, a multiplicative weighting is needed. This is simple in the case of complete destruction (e.g., in quarries) or in the case of perfect preservation (e.g., under 2 m of drift sand), but a more subtle approach is needed in the intermediate situations where sites may have been partly destroyed. The experts' weighting can provide a first estimation in this respect, and these weights can be normalized to reflect the range from optimal to minimal preservation. However, when these weights are multiplied with the site-density estimates, the resulting weights highly favor the areas where some form of protection is present (table 4.5).

---

[6] $a = \mu (((\mu (1 - \mu) / \sigma^2) - 1)$; $b = (1 - \mu) (((\mu (1 - \mu) / \sigma^2) - 1)$; from Iversen (1984)

	prior mean	conditional mean	posterior mean	posterior st.dev.	95% confidence interval	
1	9.99%	6.94%	7.06%	1.63%	4.22%	10.56%
2	12.31%	28.95%	28.40%	2.87%	22.94%	34.20%
3	3.00%	3.03%	3.03%	1.04%	1.34%	5.38%
4	1.38%	1.98%	1.84%	0.76%	0.66%	3.61%
5	12.59%	11.37%	11.41%	2.03%	7.75%	15.67%
6	12.79%	10.17%	10.25%	1.93%	6.78%	14.34%
7	7.40%	2.29%	2.57%	1.00%	0.99%	4.85%
8	9.44%	3.57%	3.82%	1.22%	1.81%	6.53%
9	8.64%	0.42%	0.80%	0.57%	0.10%	2.23%
10	4.20%	0.63%	0.96%	0.60%	0.16%	2.44%
11	0.73%	0.42%	0.53%	0.38%	0.06%	1.48%
12	9.99%	29.97%	29.16%	2.89%	23.67%	34.97%
13	4.07%	2.28%	2.45%	0.95%	0.95%	4.64%
14	3.46%	3.03%	3.08%	1.06%	1.36%	5.46%

Table 4.4a. Development of weights using Beta distributions for all units on the Ede predictive map, using all available site data (n=235). The prior Beta distribution was calculated by assuming a = 1 and taking the mean of the experts' opinions to calculate b. The conditional means were corrected for the effect of land-unit size. The resulting posterior weights should therefore not be interpreted as the percentage of sites to be found on each land unit. The 95% confidence interval is shown in the last two columns.

	prior mean	conditional mean	posterior mean	posterior st.dev.	95% confidence interval	
1	9.99%	6.94%	7.85%	1.41%	5.31%	10.83%
2	12.31%	28.95%	23.51%	2.24%	19.26%	28.03%
3	3.00%	3.03%	3.47%	0.69%	2.24%	4.94%
4	1.38%	1.98%	2.07%	0.69%	0.95%	3.61%
5	12.59%	11.37%	11.65%	1.35%	9.14%	14.42%
6	12.79%	10.17%	12.47%	1.34%	9.95%	15.22%
7	7.40%	2.29%	3.90%	1.04%	2.13%	6.17%
8	9.44%	3.57%	4.01%	1.22%	1.98%	6.71%
9	8.64%	0.42%	2.20%	0.84%	0.87%	4.12%
10	4.20%	0.63%	4.07%	0.32%	3.47%	4.72%
11	0.73%	0.42%	0.94%	0.09%	0.78%	1.12%
12	9.99%	29.97%	15.34%	1.26%	12.95%	17.88%
13	4.07%	2.28%	2.59%	0.82%	1.23%	4.42%
14	3.46%	3.03%	3.28%	0.88%	1.78%	5.22%

Table 4.4b. The same as 4.4a, but the prior Beta distribution was calculated by using the mean and standard deviations of the experts' opinions to obtain values for a and b.

	site density weight	site preservation weight	normalized site preservation weight	product of site density and preservation	normalized product
1	9.99%	4.11%	29.78%	2.98%	5.27%
2	12.31%	13.81%	100.00%	12.31%	21.78%
3	3.00%	4.83%	34.94%	1.05%	1.86%
4	1.38%	7.64%	55.30%	0.76%	1.35%
5	12.59%	5.60%	40.57%	5.11%	9.04%
6	12.79%	13.57%	98.21%	12.56%	22.22%
7	7.40%	4.45%	32.24%	2.39%	4.22%
8	9.44%	4.45%	32.24%	3.04%	5.38%
9	8.64%	13.39%	96.95%	8.38%	14.83%
10	4.20%	4.60%	33.30%	1.40%	2.48%
11	0.73%	10.94%	79.19%	0.58%	1.02%
12	9.99%	5.33%	38.61%	3.86%	6.83%
13	4.07%	6.67%	48.25%	1.96%	3.48%
14	3.46%	0.59%	4.29%	0.15%	0.26%

Table 4.5. Combined weighting of land units for site-density and site-preservation potential, both based on the mean of the experts' opinions. Preservation was normalized to reflect the assumption that the highest ranking unit equates to perfect preservation. The final weighting was again normalized to a 100% scale.

	prior mean	conditional mean	posterior mean	posterior mode	posterior st.dev.	95% confidence interval	
1	4.11%	6.25%	5.44%	4.01%	2.81%	1.35%	12.11%
2	13.81%	6.25%	7.41%	5.53%	3.77%	1.86%	16.31%
3	4.83%	6.25%	5.76%	4.26%	2.97%	1.44%	12.80%
4	7.64%	5.00%	5.65%	3.91%	3.14%	1.20%	13.19%
5	5.60%	15.00%	12.10%	10.74%	4.25%	5.10%	21.57%
6	13.57%	11.25%	11.61%	9.92%	4.61%	4.24%	22.03%
7	4.45%	5.00%	4.80%	3.31%	2.68%	1.02%	11.27%
8	4.45%	25.00%	17.61%	16.54%	4.78%	9.29%	27.90%
9	13.39%	3.75%	5.27%	3.30%	3.21%	0.91%	13.10%
10	4.60%	12.50%	9.72%	8.37%	3.74%	3.71%	18.17%
11	10.94%	2.50%	4.07%	2.12%	2.79%	0.51%	11.04%
12	5.33%	12.50%	10.21%	8.81%	3.92%	3.91%	19.06%
13	6.67%	15.00%	12.73%	11.32%	4.45%	5.37%	22.63%
14	0.59%	3.75%	1.20%	0.73%	0.75%	0.20%	3.05%

Table 4.6. Development of weights for site preservation using Beta distributions for all units on the Ede predictive map, using data from preserved sites (n=38).The prior Beta distribution was calculated by assuming a = 1 and taking the mean of the experts' opinions to calculate b. The 95% confidence interval is shown in the last two columns.

It is important to note that quantitative data on the preservation aspect can actually be obtained. For example, find spots in the national archaeological database ARCHIS, maintained by the ROB, can be registered as being intact, partially disturbed, or fully disturbed, and as such could be used to get some idea of the actual condition of sites in the various land units. However, the majority of find spots turn out not to have the relevant information registered. From the 585 find spots registered in the territory of Ede in April 2001, only 90 (15.4%) had information concerning the state of conservation at the time of discovery. Of these, 35 were completely disturbed, 45 were partially disturbed, and only 10 remained intact. For a traditional inferential model, these numbers are far too small to justify a statistical analysis. However, in a Bayesian

model it is perfectly acceptable to add this information to the expressed priors. By counting the partially disturbed sites as 50% intact, the ratio of disturbed to intact sites becomes 52/38. As was done for the site densities, a Beta distribution was also used to model both the prior and posterior belief. Table 4.6 shows the results of including the site-disturbance data.

4.5. CONCLUSIONS

The case study presented here is a first effort at incorporating multicriteria decision making and Bayesian statistics into predictive modeling. Obviously, a number of potential problems still to need be addressed:

- The 'objective' data used in this case study are not likely to be a representative sample. Bayesian statistical methods are not dependent on the strict sampling conditions to be observed with frequency statistics, but the sample should in some way reflect the target population.
- Modeling by means of Beta distributions becomes considerably more complex when multivariate modeling is undertaken. The interplay of various distributions leads to more complex mathematical models that are difficult to interpret. However, in the context of this chapter – modeling in Dutch CRM which usually deals with univariate models - this is (at least at the moment) not a major obstacle.
- The Beta distribution itself may not be the most appropiate statistical distribution; it is a form of the binomial distribution, which is applicable to very large target populations. However, in this particular case, the total number of sites in the area may not be extremely large, and a considerable proportion of it will already have been sampled. Furthermore, this sampling is done without replacement, unlike in the binomial situation. In such cases, the hypergeometric distribution is more appropriate (Davis, 1986), but its density function cannot be calculated when the size of the total population is unknown. This means that the area under consideration should be divided into equal-area sampling units (Nance, 1981). The actual size of these units will then highly influence the outcome of any statistical modeling.
- Sometimes experts are only consulted *after* a quantitative model has been constructed (e.g., Deeben *et al.*, 1997). Bayesian statistics assumes that any new information that becomes available is in the form of a sample comparable with the one used for the first calculation. A potential solution is to treat the 'conditional' experts' opinions similarly, as an independent sample, with corresponding means and standard deviations.
- The attributes chosen may be ill-considered. In theory, if an attribute is not important, it should - in the course of Bayesian model development - 'average out' and become unimportant. However, it will be harder to add a previously neglected attribute or to incorporate a 'revised' attribute (e.g., a new edition of a soil map or an improved DEM) into the model without going back to the basics of model construction.

Having said this, it is nonetheless clear that the methodological framework outlined is promising and can readily be applied to archaeological predictive modeling. In fact, apart from the use of Beta distributions, it is not using any techniques that are drastically different from earlier studies, although it does place predictive modeling in the wider context of multicriteria decision making. As such, this approach offers a more structured approach to the modeling, promises to be useful for the formalized inclusion of expert judgment in the model,

and provides an easy way to further develop the model once new data become available. The use of the Beta distributions can also tell us something about the strength or subjectivity of the model: if the posteriors are not backed up by quantitative data, then the standard deviations will become accordingly larger. The Beta distributions should also provide insight into the sample size needed to restore uncertainties to an acceptable level.

ACKNOWLEDGEMENTS

The author would like to thank the three experts who were kind enough to specify the numerical preferences for the establishment of subjective probability: Jos Deeben (ROB, Amersfoort), Eckhart Heunks (RAAP) and Huub Scholte Lubberink (RAAP). The author would also like to thank RAAP Archeologisch Adviesbureau BV for providing the opportunity, resources and time necessary to pursue this line of research.

REFERENCES

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

Aspinall, R.J., 1992. 'An inductive modeling procedure based on Bayes' Theorem for analysis of pattern in spatial data'. *International Journal of Geographical Information Systems*, 6:105-121.

Beinat, E., 1995. *Multiattribute value functions for environmental management*. Tinbergen Institute Research Series 103, Amsterdam. PhD thesis.

Bonham-Carter, G.F., 1994. *Geographic Information Systems for Geoscientists: modeling 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: modeling in the Netherlands using GIS techniques'. *World Archaeology*, 24:268-282.

Buck, C.E., W.G. Cavanagh and C.D. Litton, 1996. *Bayesian Approach to Interpreting Archaeological Data*. John Wiley and Sons, Chichester.

Burrough, P.A., 1989. 'Fuzzy mathematical methods for soil survey and land evaluation'. *Journal of Soil Science*, 40:477-492.

Burrough, P.A., R.A. MacMillan and W. van Deursen, 1992. 'Fuzzy classification methods for determining land suitability from soil profile observations and topography'. *Journal of Soil Science*, 43:193-210.

Chadwick, A.J., 1978. 'A computer simulation of Mycenaean settlement', in: Hodder, I. (ed.),*Simulation studies in archaeology*. Cambridge University Press, Cambridge, pp. 47-57.

Dalen, J. van, 1999. 'Probability modeling: a Bayesian and a geometric example', in: Gillings, M., D. Mattingley and J. van Dalen (eds.), *Geographical Information Systems and Landscape Archaeology.* The Archaeology of Mediterranean Landscape 3. Oxbow Books, Oxford, pp. 117-124.

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

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

Davis, J.C., 1986. *Statistics and Data Analysis in Geology.* Second Edition. John Wiley and Sons, New York.

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

Eastman, J.R., P.A.K. Kyem, J. Toledano and W. Jin, 1993. *GIS and Decision Making*. Explorations in Geographic Information Systems Technology, Volume 4. United Nations Institute for Training and Research (European Office), Geneva.

Finke, P., J. Hardink, J. Sevink, R. Sewuster and S. Stoddart, 1994. 'The dissection of a Bronze and Early Iron Age landscape', in: C. Malone and S. Stoddart (eds.), *Territory, Time and State. The archaeological development of the Gubbio Basin.* Cambridge University Press, Cambridge.

Goodchild, M.F., 1986. *Spatial Autocorrelation*. Catmog 47. Geo Books, Norwich.

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

Kamermans, H., 1993. *Archeologie en landevaluatie in de Agro Pontino (Lazio, Italië).* Universiteit van Amsterdam. Amsterdam. PhD Thesis.

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

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 – CAA97. Computer Applications and Quantitative Methods in Archaeology 25th Anniversary Conference, University of Birmingham.* British Archaeological Reports, International Series 750. Archaeopress, Oxford. CD-ROM.

Keeney, R.L. and H. Raiffa, 1976. *Decisions with multiple objectives: preferences and value trade-offs.* John Wiley and Sons, New York.

Kvamme, K.L., 1990. 'The fundamental principles and practice of predictive archaeological modeling', in: A. Voorrips (ed.), *Mathematics and Information Science in Archaeology: A Flexible Framework,* 257-294. Studies in Modern Archaeology Vol. 3. Holos Verlag, Bonn.

Leusen, P.M. van, 1993. 'Cartographic modeling in a cell-based GIS', in: Andresen, J., T. Madsen en I. Scollar (eds.), *Computer Applications and Quantitative Methods in Archaeology 1992.* Aarhus University Press, Aarhus, pp.105-123.

Leusen, P.M. van, 1995. 'GIS and Archaeological Resource Management: A European Agenda', in: Lock, G. and Z. Stančič (eds.), *Archaeology and Geographical Information Systems.* Taylor and Francis, London, pp. 27-41.

Leusen, P.M. van, 1996. 'Locational modeling in Dutch Archaeology', in: Maschner, H.D.G. (ed.), *New Methods, Old Problems: Geographic Information Systems in Modern Archaeological Research.* Occasional Paper no. 23. Center for Archaeological Investigations, Southern Illinois University, Carbondale, pp. 177-197.

Malczewski, J., 1999. *GIS and Multicriteria Decision Analysis.* John Wiley and Sons, New York.

Nance, J.D., 1981. 'Statistical Fact and Archaeological Faith: Two Models in Small-Sites Sampling'. *Journal of Field Archaeology*, 8:151-165.

Nijkamp, P., P. Rietveld and H. Voogd, 1990. *Multicriteria Evaluation in Physical Planning*. Contributions to Economic Analysis 185, North-Holland, Amsterdam.

Orton, C., 2000. 'A Bayesian approach to a problem of archaeological site evaluation', in: K. Lockyear, T. Sly and V. Mihailescu-Birliba (eds.), *CAA 96. Computer Applications and Quantitative Methods in Archaeology.* BAR International Series 845. Archaeopress, Oxford, pp. 1-7.

Robertson, I.G., 1999. 'Spatial and multivariate analysis, random sampling error, and analytical noise: empirical Bayesian methods at Teotihuacan, Mexico'. *American Antiquity* 64(1), 137-152.

Saaty, T., 1980. *The Analytical Hierarchy Process.* McGraw-Hill, New York.

Soonius, C.M. and L.A. Ankum, 1990. *Ede; II. Archeologische Potentiekaart.* RAAP-rapport 49. Stichting RAAP, Amsterdam.

Venneker, R.G.W., 1996. *A distributed hydrological modeling concept for alpine environments.* Vrije Universiteit, Amsterdam. PhD thesis.

Verhagen, P., M. Wansleeben and M. Van Leusen, 2000. 'Predictive modeling in the Netherlands. The prediction of archaeological values in Cultural Resource Management and academic research.', in: Hartl, O. and S. Strohschneider-Laue (eds.), *Workshop 4 Archäologie und Computer 1999*. Forschungsgesellschaft Wiener Stadarchäologie, Wien. CD-ROM.

Wansleeben, M., 1988. 'Applications of geographical information systems in archaeological research', in: Rahtz, S.P.Q. (ed.), *Computer Applications and Quantitative Methods 1988.* BAR International Series 466(ii). Tempus Reparatum, Oxford, pp. 435-451.

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

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

Wescott, K.L. and R.J. Brandon, 2000. *Practical Applications of GIS for Archaeologists. A Predictive Modeling Toolkit.* Taylor and Francis, London.

POSTSCRIPT TO CHAPTER 4

This is the only chapter I have written without external funding. RAAP was kind enough to provide me with research time to analyze the potential of multicriteria decision making and Bayesian statistics in predictive modelling. The chapter was written in the course of 2001, and was presented at a conference on archaeological predictive modelling in Argonne, Illinois, in March 2001, with the aim to make it a substantial piece of this thesis. It is part of the same volume as chapter 2 (Mehrer and Wescott, 2006). As outlined in the paper, the switch from inductive to hybrid or expert judgment modelling in Dutch CRM had already been made in 2001. I intended to find out if expert judgment could be quantified in such a way, that a statistical analysis could still be performed in a meaningful way for model development. In January 2005, a renewed attempt of combining expert judgement and quantitative techniques was made by Benjamin Ducke and Andrew Millard, using Dempster-Shafer theory and Bayesian statistics. The results of their efforts will be published in final report of the BBO Predictive Modelling project (see chapter 1).

Multicriteria decision making, though quite fashionable in the late 1990s, now seems to have lost much of its appeal to policy makers. It is not applied in archaeological decision making in the Netherlands. My colleague Daan Hallewas (ROB) opposed it, as it seemed to him the ultimate way of obscuring the decision-making process. This opinion contradicts the expressed goal of MCDM, i.e. to increase the transparency of the weighting process by using quantitative techniques. It is, however, evident that complex calculations will, in general, not increase the understanding of how a decision is arrived at. Even so, MCDM is a technique that can be applied at various levels of complexity, and it is not extremely difficult to handle at the level discussed in this chapter. I therefore believe that it can be a valuable tool when seeking quantitative answers from experts.

The hesitation on my part is more coupled to the utility of Bayesian statistics. Initially, I was quite impressed with the technique, even though it can become extremely complex when developing multivariate models. I already pointed this out in the conclusions of this chapter, but it was really made clear to me by Andrew Millard, a recognized expert on Bayesian statistics, who noted that the Beta-distribution used in the chapter should have been replaced by a Dirichlet-distribution (the conjugate distribution of the multinomial distribution; Millard, 2005) –a statistical distribution that I had never heard of before, and that was not mentioned in any of the base texts I consulted. Needless to say, this type of calculations can no longer be performed in Excel.

Furthermore, the example given in this chapter for calculating the posterior distributions is only there for purpose of illustration. In Bayesian statistical analysis, data collection follows the formulation of the prior distributions, whereas the data used in the paper were already there, and are, in all probability, not a representative sample. It is therefore impossible to conclude anything on the strength of the experts' opinions. Even though in Bayesian statistics we do not need to follow a random sampling routine, it is not good enough to wait for chance discoveries and add them to the existing sample if certain areas are systematically undersampled. This may be the case in especially the forested areas in Ede. These are mainly recreational areas and nature reserves where very few building activities are going on. Therefore, they are the least likely zones to yield new observations in the future, and they were probably not surveyed intensively in the past either. In those cases, a more pro-active form of data collection should be pursued (see also chapter 7) to obtain more accurate predictions.

However, I now believe that the real problem with Bayesian statistical reasoning is in the formulation of the prior distribution. The prior distribution can be constructed in such a way that it will either be very easy

or very difficult to reject the original belief. This completely depends on the certainty expressed by the experts about their belief, and their ability to frame their beliefs in such a way that they can be translated into statistical models. In a different context (Tol *et al.,* 2004), I tried to demonstrate that Bayesian statistics might be used to test whether a core sample containing an artefact was sufficient proof to conclude that one had hit an archaeological site. By assuming that a site exhibits an artefact density of at least 20 shards per m^2, I had to conclude that a substantial number of samples was needed to 'push' the Bayesian model into the zone where one can safely assume to have struck upon a site. In practice, two samples with artefacts are considered sufficient evidence by the archaeologists doing the survey.

Finally, Bayesian statistical methods are not capable of dealing with revisions of the original hypotheses used (see chapter 7). In the current chapter, I gave the example that a new version of a soil map cannot be incorporated in the model without rebuilding it, and the same is true for situations where the model turns out to be wrong, and needs additional parameters. This is not easy for any type of statistical model – in fact, a logistic regression model would need to be completely revised as well - but the claim, sometimes made, that Bayesian statistics constitute a superior way of doing statistical analysis, is in my view exaggerated. Given the doubts and complexities surrounding the subject, I have not further pursued its development.

ADDITIONAL REFERENCES

Mehrer, M. and K. Wescott (eds.), 2006. *GIS and Archaeological Site Location Modeling.* CRC Press, Boca Raton.

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

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

PART 2 ARCHAEOLOGICAL PROSPECTION, SAMPLING AND PREDICTIVE MODELLING

In this part, I will present the results of two research projects related to the issue of sampling by means of archaeological prospection, and its consequences for creating and testing of archaeological predictive models.

In chapter 5, a short introduction is given to the statistical background of finding the optimal method of archaeological core sampling. This paper resulted from a research project financed by the Dutch Ministry of Economic Affairs on the effectivity of core sampling for archaeological prospection, which has resulted in a more detailed publication in Dutch. The message of this chapter is that core sampling will seldom result in a complete discovery of archaeological sites. Only site types that are characterized by a relatively large size and a strong concentration of archaeological indicators will be detected by means of core sampling. For other site types, the detection probability will be much lower. For this reason, it is extremely important that before starting a prospection, it is defined what site types are looked for, in terms of expected sizes and archaeological indicators present.

In chapter 6 this conclusion is taken further by looking at alternatives to core sampling. Field survey and trial trenching also have their limitations, and the choice of the right prospection technique depends on a cost-benefit analysis of the effectivity of the various methods available and the costs involved. It is a surprising conclusion that this cost-benefit analysis is hardly ever made on the basis of probabilistic arguments. Even general guidelines on the percentage area to be covered by trial trenches are based on custom, rather than on an estimation of the probability that certain site types will be discovered. This in turn influences the quality of the archaeological site registers used for predictive modelling. The success of archaeological prospection depends on the method used. Every method has its blind spots, and only by analysing these blind spots will it become clear to what extent a prospection will give a complete image of the presence of archaeological sites. In practice, this analysis is never performed.

Chapter 7 was written as a detailed study within the project Strategic research into, and development of best practice for, predictive modelling on behalf of Dutch cultural resource management. It tries to answer the question how to test archaeological predictive models. In practice, some testing of predictive models is done, but it is not supported by statistical theory and its results are only sporadically fed back into the predictive models. Apart from that, it is not clear what should be the extent of testing in order to obtain an acceptable reliability of the predictions. The absence of statistical reasoning in everyday archaeological practice also plays an important role here. Nevertheless, even if the question of reliability is sometimes put forward, there are no tools available to provide a quantitative answer. In this chapter three aspects of testing are comprehensively discussed.

Measuring the quality of predictions

A good predictive model will result in a map where as many archaeological sites as possible will be found in a zone of ‘high probability’ that is as small as possible. The first aspect is also known as the ‘accuracy’, and the second as the ‘precision’ of the model. In the past, various methods have been developed to measure and compare these two aspects. The current research shows that the much used ‘gain’ statistic is best

suited for this, but nevertheless cannot provide a fair comparison of all situations. Gain assumes that accuracy is as important as precision. In practice a tension exists between the interests of archaeological heritage management, that will place accuracy first, and the economic and political reality, that asks for predictive models that are as precise as possible. From an archaeological point of view it is therefore recommended to define the minimal acceptable accuracy of a predictive model, and to obtain the maximum precision possible within this restriction. In this way archaeological predictive models will be easier to compare, and cannot be negotiated by trading archaeological sites for a higher precision.

Quality improvement without testing

The second aspect elaborated in chapter 7 is the option to increase the predictive power of archaeological predictive models without having to collect new, independent data. This can be done using a statistical technique known as resampling. With resampling, many sub-samples are taken from the existing data set. With these artificially created sub-samples, a better estimate of the reliability of the predictions can be made. These methods have been strongly criticized in archaeological literature, as they do not use independent data. While it is true that formal statistical testing needs independent data, this does not imply that resampling methods are useless. Rather, these methods have recently become more and more popular in statistics, as they can substantially increase the reliability of predictions within the restrictions of the available data sets. It is therefore recommended to incorporate these methods as a standard procedure for establishing the quality of archaeological predictive models.

Testing with independent data

The last subject discussed in chapter 7 is the use of new, independent data sets. What is the size of the sample necessary to obtain predictions with sufficient reliability? Two complicating factors are important in this respect. First of all, predictive maps are at the moment not presented with associated confidence limits. This seriously restricts the application of statistical testing techniques, as it is not clear what is the current and desired margin of error of the predictions. Apart from that, even when a desired confidence limit is known, it is very difficult to tell beforehand how much of an area needs to be surveyed to obtain a data set of sufficient size for testing. All available statistical methods for establishing sample size are based on the use of absolute estimates from representative samples. Neither of these conditions is fulfilled in current predictive modelling practice.

In order to obtain a good estimate of the error margin and the associated sample size needed for testing, predictive models will have to be made that estimate absolute numbers of sites on the basis of a data set that has been checked and corrected for survey biases. A preliminary survey of the research project data registered in ARCHIS (the Dutch national sites and monuments register) shows that such a data set may perhaps be obtained, but not on the basis of the characteristics registered in ARCHIS. Aspects like the extent of the surveyed zone, the number of core samples taken, and the depth of penetration, are not systematically registered. It will therefore need extra effort to collect and check the necessary data.

CHAPTER 5 Establishing optimal core sampling strategies: theory, simulation and practical implications[1]

Philip Verhagen and Adrie Tol[2]

5.1. INTRODUCTION

Archaeological core sampling is an important surveying tool in the Netherlands. It is widely used to determine the archaeological content and value of the soil record. Unfortunately, there is little documentation on the effectiveness of existing core sampling strategies for detecting and identifying specific site types.

The Senter agency of the Dutch Ministry of Economic Affairs is co-ordinating a programme of subsidized research projects aiming at promoting the use and development of technological innovations in public archaeology. Within this programme, the development of new, non-destructive ways of prospection is an important issue. In late 2001, RAAP Archeologisch Adviesbureau BV was asked to carry out a study within this programme on the effectivity of core sampling as a prospection technique. The project, that is currently near completion, has tried to gather information on this aspect and provide an assessment of the strengths and weaknesses of core sampling as a prospection technique. This paper focuses on the possibility of establishing optimal core sampling strategies for different site types, in particular through the use of simulation to predict the expected costs and benefits of each individual strategy, using the example of the excavation of Zutphen-Ooijerhoek (province of Gelderland, The Netherlands).

5.2. CORE SAMPLING: THE BASICS

Core sampling is not often used for archaeological prospection outside the Netherlands, although it is widely known as a geological survey technique. In areas where a strong accumulation of fluvial or marine sediments is found, core sampling is the only technique available that will provide a quick and cheap assessment of the local stratigraphy. Core sampling is still largely performed by means of manual labour, even though mechanical alternatives are currently being developed. Two basic types of equipment can be used. The auger has a diameter of 7 cm (sometimes 15 cm). It is screwed into the ground and takes small cores per sample (about 15 cm long). It is best suited for dry and sandy soils, and is not frequently used at depths below 2 meters. The gouge has a standard diameter of 3 or 5 cm and is driven with force into humid clayey soils or peat. The core obtained is 1 meter long. This type of core sampling can reach depths of 7 meters or even more, by extending the gouge with metal rods.

Given the fact that large areas of the Netherlands are covered with Holocene fluvial and marine sediments, it is not surprising that core sampling is also considered an appropriate tool for archaeological prospection. In many areas, there is no other way to obtain sufficient information on the (possible) presence of archaeological remains. In fact, its use has resulted in the discovery of some very important archaeological sites, like those found in the alignment of the Betuweroute-railway, that runs straight through the river basin of Rhine and Meuse (Asmussen and Exaltus, 1993; Asmussen, 1994).

---

[1] This paper also appeared in A. Fischer Ausserer, W. Börner, M. Goriany and L. Karlhuber-Vöckl (eds.), 2004*: [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. Archaeopress, Oxford, pp. 416-419.

[2] RAAP Archeologisch Adviesbureau BV, Amsterdam, The Netherlands. The text for this paper was prepared by me, the research was done in close collaboration with Adrie Tol.

5.3. STATISTICAL BACKGROUND

The probability of discovering an archaeological site by means of any method of 'small unit sampling' (other possible methods are test pit sampling and machine trenching) is given by the following equation:

$$P = I \bullet D$$

where

P = discovery probability;
I = intersection probability; and
D = detection probability.

The intersection probability describes the relationship between the size of the object to be found and the distance between the sampling points. It can be determined using the following equation (Drew, 1979):

$$I = A / (i \bullet s)$$

where

A = the area of the object;
i = the distance between the sampling points in a row; and
s = the distance between the rows.

This equation does not take into account the form and position of the objects. Krakker *et al.* (1983) have demonstrated that the optimal layout for a sampling grid is an equilateral triangular grid. In this case, the distance between rows *s* equals $\frac{1}{2} i \sqrt{3}$. For a standard core sampling survey, with sampling points every 50 meters, this equates to a distance between rows of 43.3 meters. The maximum diameter of a circular object that can be missed by such a grid layout is equal to $s + (i^2 / 4s)$, or 57.73 meters in the case of a standard grid (Kintigh, 1988).

For elongated (elliptical) objects, the mean intersection probability is the same as for circular objects, but the probability distribution is different, and they may therefore slip through the net more easily (Gilbert, 1987). However, when looking for elliptical objects, it is not necessarily useful to change the layout of the grid. Drew (1979) stated on theoretical grounds that using a rhomboid instead of an equilateral triangular grid is only effective when the orientation of the objects is more or less known. However, simulations carried out by ourselves show that there is a small positive effect of finding extremely elongated objects by using a rhomboid grid, even when the orientations are not known.

The detection probability for archaeological artefacts is given by the following equation (Stone, 1981; Krakker *et al.* 1983):

$$D = 1 - e^{-A \bullet d \bullet W}$$

where

e = the base of natural logarithms (2.711828);
A = the area of the sampling unit;
d = the density of artefacts per area unit; and
W = the observation probability.

This equation describes a Poisson-distribution, that is appropriate for rare objects that are not very likely to be encountered in a sample. Artefact density determines whether a site may be detected or not, but the

observation technique chosen determines whether an artefact will actually be observed. Very little data are available on the effects of sieving versus visual inspection, or of choosing a different sieving mesh. Groenewoudt (1994) showed that about 75% of the flints found at the site of the Ittersumerbroek excavation were smaller than 4 mm, so choosing a smaller sieving mesh may drastically increase the amount of observed artefacts.

Very few data are available on the actual artefact densities encountered on archaeological sites in the Netherlands. Mean artefact density estimates are given by some authors. Groenewoudt (1994) for example estimated the mean artefact densities for Iron Age and Roman settlements at more than 120 shards per m^2, an estimate obtained by extrapolating data from core samples. It should be noted that the actual detection probability of such a density is not very high when using a standard 7 cm auger (about 37%). For a selection of 79 Stone Age sites from NW Europe (kindly put at our disposal by dr. Willem-Jan Hogestijn) the mean artefact density is 140.4 per m^2, but 70.9% of these sites have densities below 50 per m^2. In the recent excavation of the Mesolithic site of the Hoge Vaart by Hogestijn and Peeters (2001), mean flint densities of only 18 and 16 per m^2 were registered when sieving with a 2 mm mesh. Groenewoudt (2002) also mentions an example of a site with a mean density of only 6.4 artefacts per m^2 (sieved with a 4 mm mesh); the site actually contained two house plans.

The observation method used is obviously very important in this respect. Core sampling is based on very small sampling units, the samples are usually thoroughly described, and the soil is sieved with a 1 mm mesh to obtain as many artefacts and other archaeological indicators as possible. Archaeological features are not usually recognized in core samples. During excavations, or even in machine trenching surveys, the features are of primary concern, and artefacts are usually only collected and described if they have diagnostic value. Given the already enormous amounts of artefacts collected in this way (e.g. almost 40,000 in the Malburg excavation; Oudhof *et al.*, 2000), it is very understandable that a full count of all artefacts present per feature or quadrat is not performed. However, this implies that it is impossible to obtain reliable data on the spatial distribution of artefact densities.

Only a few examples could be found of sites that had been consistently sieved for artefacts in quadrats, and all of these concerned small excavated areas with relatively low artefact densities. Simulations performed on these data showed that these sites will be very difficult to discover by means of standard core sampling survey.

5.4. ESTABLISHING AN OPTIMAL CORE SAMPLING STRATEGY: THE CASE OF ZUTPHEN-OOIJERHOEK

The Mesolithic site of Zutphen-Ooijerhoek[3], for example, was sieved with a 3 mm mesh in 50 by 50 cm quadrats. The resulting flint counts ranged from 0 to 179, resulting in a mean artefact density of 66 per m^2, on a total excavated area of 246.75 m^2. A strong clustering of the flints was evident; in about two-thirds of the excavated area, the artefact density was below average. For purposes of comparison, the centre of the site was analysed separately from the periphery (see table 5.1). The probability of finding the site using standard core sampling strategies was approached by simulating 1,000 hypothetical surveys of the site, using different parameters for grid size and sample diameter. In this way, the costs and benefits of each strategy can be compared. The probabilities given in table 5.1 should however not be seen as real probabilities of finding the site, as the effect of the observation method chosen has not been incorporated in the simulation runs.

[3] The data of the Zutphen-Ooijershoek excavation were kindly put at our disposal by drs. Jos Deeben, Rijksdienst voor het Oudheidkundig Bodemonderzoek, Amersfoort.

ZUTPHEN-OOIJERHOEK	centre	periphery	total	cost factor
mean artefact density per m^2	*165.84*	*21.04*	*66.08*	
area in m^2	*76.75*	*170.00*	*246.75*	
discovery probability 7 cm auger				
40 x 50 m	*1.6%*	*0.1%*	*3.1%*	*1*
20 x 25 m	*6.2%*	*2.8%*	*7.7%*	*4*
10 x 12.5 m	*22.4%*	*9.2%*	*33.6%*	*16*
6 x 6.25 m	*64.8%*	*28.5%*	*73.4%*	*64*
discovery probability 15 cm auger				
40 x 50 m	*3.6%*	*2.3%*	*5.3%*	*2*
20 x 25 m	*11.1%*	*9.4%*	*19.1%*	*8*
10 x 12.5 m	*43.5%*	*34.7%*	*63.8%*	*32*

Table 5.1. Comparison of the costs of different core sampling strategies for Zutphen-Ooijerhoek, based on simulation results. The centre of the site is the area where artefact density is above average. An increase in grid density means a four-fold increase in number of samples, an increase in auger diameter implies a two-fold increase in time needed to take, sieve and describe a sample.

It turns out that for this particular site, increasing the sample volume is a more cost-effective strategy than applying a denser sampling grid. However, it should be taken into account that taking a larger sample volume is a course of action that can only be applied once, as augers with a larger diameter than 15 cm are not available.

5.5. CONCLUSIONS

The results of the simulations, as well as theoretical considerations, point to the conclusion that core sampling is not a very effective technique to discover small archaeological sites when they have a low density of artefacts. Even without the availability of much representative data on artefact densities from excavations, it can be suspected that especially Stone Age (and other briefly occupied) sites run this risk, as well as off-site phenomena.

However, artefacts are not the only category of indicators looked for and registered in a core sampling survey. In fact, three classes of indicators are registered. The first of these are non-archaeological, like soil type and lithology which can serve as predictors of possible site locations. Secondly, there are (semi)-archaeological indicators with a higher detection probability than artefacts, like charcoal. Even if these indicators are not hard evidence of an archaeological site in the sense that artefacts are, they are almost certainly evidence of human occupation very near to the sampled location. Only in third instance 'real' archaeological indicators come into play, as the final corroboration that we are dealing with an archaeological site. It is only when geomorphological, pedological and archaeological 'predictors' are either absent or too small in size for detection in a standard core sampling survey that low density artefact scatters are likely to escape detection, as there will be no apparent reason to 'zoom in' on a specific location.

The absence of reliable data on the density and spatial distribution of indicators for different types of archaeological sites in the Netherlands makes it difficult to design site-specific prospection strategies. These data can only be obtained by registering the same data in excavations as during core sampling, and will need to be collected in a systematic way during future excavations and trenching campaigns. However, at the moment this is not happening in Dutch public archaeology, also because core sampling and excavation are often carried out by different commercial parties, that may not perceive the mutual benefit that can be obtained from

investing time and money in this type of work. It is therefore hoped that the current project will provide the necessary impetus to actually start the comparative research needed for further improvement of archaeological prospection strategies in the Netherlands.

REFERENCES

Asmussen, P.S.G., 1994. *Archeologische Begeleiding Betuweroute. Deel C: Waardering van de vindplaatsen.* RAAP-rapport 86. Stichting RAAP, Amsterdam.

Asmussen, P.S.G. and R.P. Exaltus, 1993. *Archeologische Begeleiding Betuweroute. Deel B: Inventarisatie. Deel C (gedeeltelijk): Waardering.* RAAP-rapport 76. Stichting RAAP, Amsterdam.

Drew, L.J., 1979. 'Pattern Drilling Exploration: Optimum Pattern Types and Hole Spacings When Searching for Elliptical Shaped Targets'. *Mathematical Geology* 11:223-254.

Gilbert, R.O., 1987. *Statistical Methods for Environmental Pollution Monitoring.* Van Nostrand Reinhold Company, New York.

Groenewoudt, B.J., 1994. *Prospectie, waardering en selectie van archeologische vindplaatsen: een beleidsgerichte verkenning van middelen en mogelijkheden.* Nederlandse Archeologische Rapporten 17. Rijksdienst voor het Oudheidkundig Bodemonderzoek, Amersfoort.

Groenewoudt, B.J., 2002. 'Sieving Plaggen Soils; extracting Historical Information from a Man-made Soil'. *Berichten van de Rijksdienst voor het Oudheidkundig Bodemonderzoek* 45: 125-154.

Hogestijn, J.W.H. and J.H.M. Peeters, 2001. *De mesolithische en vroeg-neolithische vindplaats Hoge Vaart-A27 (Flevoland).* Rapportages Archeologische Monumentenzorg 79. Rijksdienst voor het Oudheidkundig Bodemonderzoek, Amersfoort.

Kintigh, K.W., 1988. 'The effectiveness of subsurface testing: a simulation approach'. *American Antiquity* 53: 686-707.

Krakker, J.J., M.J. Shott and P.D. Welch, 1983. 'Design and evaluation of shovel-test sampling in regional archaeological survey'. *Journal of Field Archaeology* 10: 469-480.

Oudhof, J.W.M., J. Dijkstra and A.A.A. Verhoeven, 2000 (eds.). *Archeologie in de Betuweroute: 'Huis Malburg' van spoor tot spoor: een middeleeuwse nederzetting in Kerk-Avezaath.* Rapportage Archeologische Monumentenzorg 81. Rijksdienst voor het Oudheidkundig Bodemonderzoek, Amersfoort.

Stone, G.D., 1981. 'On artifact density and shovel probes'. *Current Anthropology* 22: 182-183.

CHAPTER 6 Prospection strategies and archaeological predictive modelling[1]

6.1. INTRODUCTION

A key problem in predictive modelling is the availability of representative archaeological input data that can be used either as input to an inductive predictive model, or as a test set for an independent check of the model. Almost all available archaeological data sets are biased in one way or another to specific site types or regions. Some of this bias originates as a result of the archaeological prospection techniques used for discovering sites.

The aim of archaeological survey is to establish without any doubt the presence of archaeological sites. Subsidiary goals might be defined for a survey, like the determination of the exact location of the site, its type and dating, the layout of the site and even the conditions of the buried artefacts (see e.g. Hey and Lacey, 2001). For predictive modelling however, it suffices to obtain evidence of the presence or absence of an archaeological site, at a location that is as precise as possible. Dating and typology of a site are desirable properties to be known, but if they are not available, a non-specific predictive model might still be constructed.

The definition of an archaeological site on the basis of survey data however is problematic in itself. Tainter (1983) provides a useful working definition, that is cited by Zeidler (1995), in which the criterion for defining an archaeological site is the presence of at least two different artefacts in close proximity, or other evidence of purposive behaviour, such as archaeological features or architectural remains. Two different objects is the minimal archaeological manifestation which will consistently reflect purposive behaviour, whereas a single object cannot differentiate accidental loss. This definition does not take into account the possibility that the artefacts may be encountered *ex situ*, but it serves well as a minimal standard.

However, if only one or even no artefacts are found during survey, one cannot be certain that there is no site. The degree of confidence for establishing the absence of an archaeological site is highly dependent on the survey method chosen, and surveys may therefore underestimate the actual number of sites in an area by varying degrees. This has severe consequences, both for the curators who want to be as certain as possible that all sites in a region have been found during survey, as well as for predictive modellers, who depend on representative site samples to develop their models.

6.2. PROSPECTION STRATEGIES

The most frequently used prospection strategies for discovering archaeological sites are (non-intrusive) field survey, and intrusive methods like machine trenching, test pitting and core sampling[2]. Aerial photography (non-intrusive) and geophysical survey (intrusive, but not destructive) are also widely used for

[1] This paper also appeared in M. van Leusen & H. Kamermans (eds.), 2005: *Predictive Modelling for Archaeological Heritage Management: A research agenda*. Nederlandse Archeologische Rapporten 29. Rijksdienst voor het Oudheidkundig Bodemonderzoek, Amersfoort, pp. 109-122.

[2] the term 'intrusive' is adopted from Hey & Lacey (2001); these methods are also referred to as 'invasive' (Orton, 2000b) or as subsurface testing methods

archaeological prospection, but cannot be considered as tools that will result in the discovery of archaeological sites, as they will produce neither artefacts nor features. They rather indicate the presence of potential site locations[3], and can therefore serve as the basis for -very localized- predictive models. These potential site locations will then still have to be checked for their archaeological significance by some other form of (preferably intrusive) survey.

FIELD SURVEY

A cost-effective method of archaeological prospection is field survey[4]. Site locations are detected by walking across parcels of cultivated land and recording all archaeological materials found[5]. It is however a method with a number of drawbacks. The potential of field walking for discovering archaeological sites is limited by the fact that field walking will only reveal information on archaeological remains that are within a depth of approximately 50 cm (the plough zone). In areas without cultivated land (grassland, forests), its intrusive capacities are obviously much lower. Furthermore, even in zones that are regularly ploughed, the results of field survey are highly dependent on the time of year and the weather conditions.

Field walking is not necessarily the preferred method of survey in Cultural Resource Management (CRM). Hey and Lacey (2001) state that field walking is a good method to indicate site presence and dating, but they rate it below machine trenching because of its inability to determine the exact location and layout of a site. In the Netherlands, field walking is not the most frequently used type of survey; core sampling is usually preferred, because of the abundance of grassland and the geomorphological conditions. The Florida Department of Transportation (2001) does not even mention field walking as a method to be applied for CRM survey; instead, test pitting is the only technique allowed. The demands posed by CRM-archaeologists on survey seem to favour methods that do not have the drawback of only providing information on the topsoil. The greatest advantage of field survey, apart from the low investment in time and material, is its potential to obtain a complete surface coverage[6] of the area surveyed. Therefore, it is still regularly applied in situations where costs are important, and larger areas are concerned.

INTRUSIVE TECHNIQUES

Intrusive prospection is not hampered by the visibility problems encountered in field survey. In contrast to field survey however, all intrusive prospection methods only sample a very small portion of the surveyed area, as full coverage would result in complete excavation. Furthermore, the actual depth of penetration may vary widely between surveys.

Machine trenching[7] is currently the most widely used intrusive prospection method in European CRM archaeology. It is relatively cheap, as large areas may be uncovered in a short period of time.

---

[3] an exception can be made for situations where features visible from the air are so distinctive as to be interpretable as specific archaeological sites (see e.g. Wilson, 2000); this is also true for geophysical prospection

[4] also known as field walking or pedestrian survey

[5] at least, in areas where artifacts are relatively sparse; in areas with very high artifact densities, this may be impossible; furtherrnore, artifacts that are considered ‘off-site’ are often not recorded

[6] this is also called intensive or systematic survey.

[7] also known as *backhoe trenching* in the United States

Test pit sampling[8] is slower per area unit than machine trenching, as it involves digging by hand and systematic sieving[9] of the artefacts, whereas machine trenching (at least in the United Kingdom) is, for financial reasons, usually only accompanied by a visual inspection of the trenched area, and artefacts from the 'spoil heaps' are not systematically collected. The area uncovered by means of test pits is therefore considerably smaller than with machine trenching.

Core sampling[10] has the distinct disadvantage of using extremely small sampling units. A typical test pit ranges in size from 30 x 30 cm to 1 x 1 m, whereas standard manual coring equipment uses a 7 cm diameter auger (for use in sandy soils) or a 3 cm gouge (for use in clayey soils or peat). Larger diameters are available, but this type of equipment quickly becomes cumbersome and is not used in standard surveys, although it is sometimes applied in situations where a larger soil sample is desired. The small volume of core samples taken can easily result in the non-discovery of archaeological remains. However, as it is the only method capable of easily penetrating the soil at depths of more than about 1.5 metres, it is an indispensable tool for prospection in areas where a strong accumulation of peat or sediments is found, and where archaeological sites may be buried at considerable depths.

6.3. CONTROLLING SURVEY BIASES

The probability that a site will be discovered (or not) is determined by the following factors:

- survey intensity
- spatial layout of survey
- size of sampling units
- the visibility of archaeological remains at the surface: obscuring of artefacts by vegetation cover, stone cover, weather conditions, agricultural practices and/or geomorphological conditions
- observation methods (sieving)
- recording practice (definition of sites, specific interest for certain site types, experience of survey crew)

All these factors can play a role in creating biases in survey reports. In order to be able to compare survey results, it is necessary to control as many of these biases as possible. The remainder of this paper will discuss the quantitative effects of survey intensity and observation methods on site discovery, and how to control for the resulting biases. Recording practice is outside the scope of this paper, as it will be much more difficult to quantify its effects, although it may evidently influence the choice for a particular survey technique.

Krakker *et al.* (1983) introduced the concepts of intersection probability and detection probability for modelling the probability of site discovery given a certain survey intensity. Basically, the probability that an archaeological site will be discovered is the product of the probability that it is intersected by field walking lines, trenches or boreholes, and the probability that the artefacts or features in the site will actually be detected. When the detection probability is equal to 1, then intersecting the site is sufficient for discovery. However, in most cases this probability is much lower.

---

[8] also known as *shovel testing* in the United States; it is for example popular in Scandinavia, in forested environments where machines can't come

[9] also known as *screening* in the United States

[10] generally referred to as *augering* or *coring* in the United States; augering is sometimes reserved for mechanical equipment, and coring for hand-powered tools

6.4. INTERSECTION PROBABILITY

The intersection probability of a site of a certain size is directly related to the spacing between field walking lines (also known as *crew spacing*), trenches or test pit and boreholes. Davis (1986:289-295) presents the equations for determining intersection probabilities of elliptical targets by parallel lines (see also Sundstrom, 1993). Drew (1979) gives the equation to be used for regular point sampling, that is applicable to test pits and boreholes and has been introduced in archaeology by Krakker *et al.* (1983). For trenches (basically a form of regular polygon sampling), no comparable equations have been published, but intersection probabilities might presumably be obtained by treating the trenches as discontinuous transects, and increasing the potential target's size by the trenches' width (see Orton, 2000a).

The intersection probability equations are presented as measures to be calculated for site sizes smaller than the spacing between sampling points, which will always result in a number smaller than 1. In fact, the number calculated by the equations is the mean of the probability distribution of hits inside a site of a certain size, given a certain spacing between sampling points or lines, and it is equally applicable for sites larger than the line or point spacing (see also Sundstrom, 1993). Only for circular targets and line walking an intersection probability of 1 implies that the site will never be missed. When using an equilateral triangular point sampling layout, the "probability" at which a circular site will always be intersected is 1.21. For an elliptical site with a ratio of major to minor axis of 2:1, this figure equals 1.67 (Visual Sample Plan 1.0; Gilbert *et al.*, 2001). Even though the actual shape and size of a site will never be known beforehand, probability distributions can be calculated for various survey layouts and target shapes and sizes by means of computer simulation.

The distance between transects or sampling points that is chosen should therefore depend on the expected site sizes and shapes. It is however not clear whether these considerations play an important role in choosing a particular spacing. Field walking distances typically seem to be in the range of 20 to 40 metres. Zeidler (1995) reports an analysis of 62 pedestrian surveys throughout the United States, dating from 1975 to 1990, and shows that mean spacings are in the order of 17 to 50 metres, the closer spacings generally being applied in areas with denser vegetation. In Attema *et al.* (2002:133-143) however, a trend is observed towards intenser survey with walker distances of only 5 to 10 metres. As this figure relates to surveys conducted by academic institutions in the Mediterranean, it is not clear if this trend is also observable in CRM-surveys and in other regions. The state of Mississippi requires a field walking distance of 15 to 30 meters for CRM surveys (Sims, 2001), and the state of Georgia requires a maximum interval between walkers of 30 meters (Georgia Council of Professional Archaeologists, 2001). Hey and Lacey (2001) report line walking distances of 20 meters in the United Kingdom, but the internal guidelines defined by RAAP for field survey in the Netherlands specify a walking distance of 5 or 10 metres. These distances refer to the diameters of the largest circular site to be missed and it seems there is a silent agreement that most sites will actually appear as roughly circular artefact scatters in ploughed fields. However, it seems highly improbable that the number of small sites in the Netherlands or the Mediterranean is considerably larger than in the United States or the United Kingdom, so an element of risk assessment will also be present in the choice for a particular line spacing.

As a consequence, the 'full area coverage' that can be obtained by field walking is relative to the line spacing chosen. However, even full area coverage, whatever its real value in terms of site discovery, is not always within reach of the available budget. Samples of the study area can then be walked. Transect survey, taking only a sample of possible field walking lines, has been applied in the Agro Pontino Survey project (Voorrips *et al.*, 1991). It was intended as a means to obtain a representative sample of archaeological remains that could then be used for extrapolation. A similar approach is found in quadrat survey, in which the survey

area is divided into quadrats of equal size, that can then be sampled according to predefined statistical confidence limits and/or budget (Nance, 1981; 1983). Quadrat survey is more time consuming than transect survey, as the individual quadrats to be sampled will be more difficult to locate in the field than the starting point of a transect. Orton (2000b) states that sampling of irregularly shaped units does not have to be problematic, so instead of selecting quadrats or transects, a selection of parcels may be equally suitable, provided some correction is made for the different sizes of the units. These approaches can be used to estimate the total number of sites in an area, and as such may all be well suited for input in an predictive model.

For core sampling and test pit sampling (which can be thought of as equivalent to point sampling), it is not only the shape and size of the expected target that determine the best survey layout, the shape of the sampling grid itself is important as well. The optimal point sampling layout is proved to be an equilateral triangular grid by Krakker *et al.* (1983). The equilateral triangular layout has been the favoured survey design for core sampling in the Netherlands for over 15 years, even though it is not mentioned as such in the handbook of Quality Norms for Dutch Archaeology (KNA; College voor de Archeologische Kwaliteit, 2001). The KNA does specify that distances between boreholes should depend on the smallest object that can be expected. In practice, standard surveys commonly apply a triangular grid of 40 x 50 meters, and a 20 x 25 meter grid is regularly used in areas where smaller sites are expected.

The American state guidelines do not specify the preferred layout of test pits, only the distance between pits. Only in Lousiana, a (presumably rectangular) grid of 30 x 30 meters is given as a guideline for high potential zones, and 50 x 50 meters for low potential zones (Louisiana Division of Archaeology, 1999). Other states only specify the distance between test pits, which effectively means that rectangular grids are applied there as well. The states of Mississippi and Georgia require a distance of 30 meters (Sims, 2001; Georgia Council of Professional Archaeologists, 2001), Virginia takes 50 feet (approx. 15 meters; Virginia Department of Historic Resources, 2001), and in Florida the distance depends on the zone on the predictive map used (25 meter in high potential, 50 meter in medium potential and 100 meter in low potential; Florida Department of Transportation, 2001). The guidelines consulted do not specify why a particular distance is chosen, but they may well be based on expected smallest site diameters, like in the case of field walking.

For trenching, spacing between trenches does not seem to play a major role in designing the survey. A layout of staggered trenches is often used, as it is proved to be more effective for intersection of sites (Hey and Lacey, 2001); typical trench sizes are 30 x 2 meters. However, the main concern in setting up a trenching survey is the percentage of the area uncovered. Of course, increasing the percentage covered while maintaining the same trench size implies that distances between trenches will be smaller. The total area covered by a typical trenching campaign seems to be about 2% in the United Kingdom, but this figure has no particular archaeological or statistical reasoning behind it (Champion *et al.*, 1995; Orton, 2000b). Hey and Lacey (2001:49-51) conclude on the basis of simulations that a 3 to 5% coverage is actually needed in order to provide an assessment of most sites that is sufficient to meet planning requirements. At about 10% coverage no significant gain can be expected from increasing the sample area. However, Orton (2000b:120-121) makes clear that from a statistical perspective the most important parameter involved is not the proportion of the terrain sampled but the absolute size of the sample. So, the best approach would seem to determine a trench spacing that will allow for the intersection of the site of interest, and then maximize the number of trenches - their minimum size being dependent on the detection probability of the features. The experiments reported in Champion *et al.* (1995) suggest that a strategy of rather large test pits may be the most efficient.

6.5. SURVEY INTENSITY AND TESTING OF PREDICTIVE MODELS

One point is worth noticing here from the point of view of predictive modelling: the fact that different distances are sometimes recommended for test pit sampling in high, medium and low potential zones. The spacing between test pits is in those cases dependent on the expected density of sites, not on the expected size of the sites. There is, after all, no apparent reason why sites in high potential zones should be smaller than in low potential zones. When using a 25 meter test pit interval, a maximum (circular) site radius of 17.65 meters may be missed. This same diameter has an intersection probability of only 39.15% at a 50 meter test pit interval, and 9.79% at a 100 meter interval (Visual Sample Plan 1.0; Gilbert *et al.*, 2001). Smaller test pit intervals will therefore result in the discovery of more small sites, so effectively a larger percentage of the site population will be discovered. This practice is the consequence of an economic consideration. Halving the distance between test pits implies a four-fold increase in sampling points, and therefore in costs. In order to reduce the field survey costs, missing a few sites in a low potential zone may be an acceptable solution, certainly when compared to missing lots of sites in a high potential zone. However, it displays a degree of trust in the predictive models used that may not be warranted, and it is probably not based on a cost-benefit analysis of the risk of missing certain categories of sites.

From the point of view of testing of predictive models, the survey practice of oversampling high potential zones will lead to data sets that are biased to those zones. This means that the confidence intervals of site density estimations in high potential zones will be much narrower than for low potential zones. There are two parameters involved in order to determine what sample size is needed to obtain site density estimates within a certain confidence limit. These are the expected site density (taken from the predictive model), and the desired confidence interval for our estimate. As the outcome of the samples can result in positive or negative evidence of an archaeological site, we can reformulate the problem in terms of proportions, and obtain the sample size required using the equation given in Shennan (1988:312-313). The proportions estimated should in this case be thought of as the percentage of an area taken up by 'significant archaeological remains' (Orton, 2000a), but it might equally well be the area covered by a geomorphological unit that is usually indicative of site presence. In cases where no predictive model is applied, all outcomes are equally likely, and the sample size needed will only depend on the width of confidence interval chosen. If a 95% confidence interval is desired whose width is 10% , then we would need 384 samples per unit regardless of the size of the unit involved[11].

When we do use a predictive model, and we are willing to postulate an expected site density for a zone (or in Bayesian statistical terms, formulate a prior belief), fewer samples will suffice when the desired confidence interval does not change. However, our primary interest may not be a site density estimate, but a decision rule that we can use to distinguish between zones. For purpose of illustration we can think of the following situation. The threshold between low and medium potential is placed at 10%, and between medium and high at 40%. In order to know whether we are in one zone or an other, we will have to know if the estimate obtained from our sample is below or above the threshold value. For this, the equation given in Orton (2000b:217) can be used, which tells us that to be 95% certain that we are dealing with a low potential zone we will need 28.4 negative samples, and no positive ones[12]. Similarly, for deciding that we are inside a medium

[11] as long as the true proportion lies approximately between 0.3 and 0.7 (Shennan, 1988:312)

[12] this equates to postulating an expected mean proportion of 3.4% with a 95% confidence interval width of 6.6%; with small proportions, the upper tail of the confidence interval is much larger than the lower tail

instead of a high potential zone, it suffices to have 5.9 negative samples. These figures become higher when smaller proportions are concerned, because narrower confidence intervals are then desired.

The whole procedure of testing predictive maps in this way becomes a matter of developing a Bayesian sampling framework. The beta distribution may serve as an approximation of the actual distribution of succesful and unsuccessful trials (Orton, 2000a). This is allowed as long as the chance of a successful trial is not dramatically low, in which case a gamma distribution might be more appropriate. In this way, it is relatively easy to perform predictive model testing on a local scale, using a moderate number of samples. However, it does imply that predictive models need to be specified in terms of area proportions occupied by significant archaeological remains, rather than by site density estimates, as it is the size of the site that determines whether it will be hit, not the number of sites. From a risk management perspective, this may even be a preferable way of defining potential zones, as the costs associated with excavating archaeological sites are of course strongly related to the size of the sites.

However, before using survey data in this way, it is necessary to define what constitutes a successful or unsuccessful trial, as this is dependent on the detection probability of the archaeological phenomena looked for.

6.6. DETECTION PROBABILITY

Detection probability determines whether an intersected object will actually be observed. In the case of clearly visible occupation layers, there will be no problem observing the phenomenon concerned, even when using core samples. Archaeological features are also easy to observe, but as they are relatively small and widely spaced, and need to be viewed in context from above or in a section in order to be recognized with certainty, small unit interventions are not very well suited for detecting features, and trenches are to be preferred. When small unit sampling is chosen as a survey method, the probability of finding an archaeological site is highly dependent on the artefact density.

The probability of encountering an artefact in a small size sample unit is given by an exponential distribution (Stone, 1981). This distribution allows us to calculate the probability of detecting a certain artefact density, given a particular sample unit size. The theory presented in Krakker *et al.* (1983) assumes that within a site the artefacts are randomly distributed, but this is not very likely to be true. Artefacts will appear as concentrations on a certain location, and the density will decrease from the site center to zero at the edge, or to a background noise of 'off-site' scatter. This decay will occur at varying distances, and will follow varying decay curves. Kintigh (1988) attempted to model different spatial artefact distributions, but it is not clear whether these distributions bear a great similarity to reality. Simulations carried out by Tol *et al.* (in prep.) show that the actual detection probability of artefacts of the Mesolithic site of Zutphen-Ooijerhoek can be considerably lower than the random distribution suggests because of the high degree of artefact clustering.

Two factors can be manipulated in order to increase detection probability: the sampling unit size, and the observation method applied. As detection probability is directly related to the density of the phenomena to be observed (in most cases the artefacts), it is necessary to choose the most effective combination of sampling unit size and observation method for money. In field survey, the observation window can be partly obscured by vegetation or stone cover, and in that way the amount of artefacts that can be observed is decreased. There will however be a point at which visibility is so much reduced, that intrusive methods will have to be used in order to make useful observations. The point at which this becomes necessary is not only dependent on the

percentage of visible terrain (as is often suggested in American state survey guidelines), but also on the expected artefact density at the surface.

In general however, it can be stated that the observation windows for both field survey and machine trenching are sufficiently large to permit the observation of artefacts without sieving. However, for test pitting and core sampling, the small sample unit size is much more problematic in this respect. Typical test pit sizes in the United States are 30 x 30 cm (0.09 m^2), and for core sampling diameters of 7 cm (0.004 m^2) or 3 cm (0.0007 m^2) are used. Sieving is therefore regularly performed for test pits (using a ¼ inch - 6.35 mm - mesh) and core samples (using different mesh sizes of 1, 2, 3 or 4 mm) in order to counteract the effect of small test unit sizes. The lack of reliable data from completely excavated archaeological sites is however a major obstacle for modelling artefact distributions. As sieving is not regularly performed in excavations or even trenches, artefact counts registered are far below the densities that can be observed with systematic sieving. And even when sieving is performed, the influence of using different mesh sizes is not sufficiently known (Verhagen and Tol, in press). It has been established by Groenewoudt (1994) that about 75% of the flints found at the Ittersumerbroek excavation are smaller than 4 mm, so decreasing the mesh size may drastically increase the observable density of artefacts.

6.7. LARGE OR SMALL INTERVENTIONS?

It is interesting to observe the different approaches to subsurface survey in the United States, United Kingdom and the Netherlands. Whereas in the United States and the Netherlands survey is primarily done by means of small unit sampling, in the United Kingdom machine trenching is the preferred method. Both core sampling and test pit sampling have been regarded with suspicion by the archaeological community as survey techniques because of the small size of the test units involved. Shott (1989) even stated that test pit sampling is a survey method "whose time has come and hopefully gone". As is clear from the current American state survey guidelines, this wish has not come true. However, augering is not generally accepted in the United States as a survey method to be used in a reconnaisance survey. The Florida Department of Transport (2001) even states that "the use of soil augers as the primary means of subsurface testing is unacceptable". Champion *et al.* (1995:54) conclude on the basis of simulations that test pit sampling performs much better for site discovery than machine trenching, because of the more detailed inspection of the finds. Hey and Lacey's (2001) opposite conclusion that machine trenching performs much better than test pitting is based on a comparison without sieving. As test pits will not easily lead to the identification of features, this is not surprising, but it should be noted that the comparison they present is not completely fair, as it is based on different area coverages (sample fractions of 2-10% for the trenches compared to 0.25% for the test pits).

Even though the scientific opinion seems to favour machine trenching, the choice for a particular survey technique should ideally depend on the archaeological and geomorphological circumstances. In order to choose a technique, one should have prior information on the archaeology and geomorphology in an area, concerning site size, depth and density of artefacts and/or features. Unfortunately, this information is precisely what we are looking for and can not be specified with certainty beforehand. Formal or informal predictive models may therefore play a very important role in choosing a particular survey strategy. Moreover, certain techniques are only preferred because of the lower costs involved; the resulting loss of information may be acceptable when it means saving money on archaeological investigations. It is however only by means of a cost-benefit analysis that an assessment can be made of the costs of a particular survey method compared to its results. In practice however, these are not customarily performed or consulted, even though a number have

been published (e.g. McManamon, 1984; Kintigh, 1988; Zeidler, 1995; Champion *et al.*, 1995; Hey and Lacey, 2001).

It may well be that the choice for a particular technique is also the result of the way in which CRM-archaeology is organized. Both in the United States and the Netherlands, reconnaisance (Phase I) and evaluation (Phase II) surveys[13] are seen as separate stages, that do not need to be carried out consecutively. Both countries also apply predictive models in order to plan their Phase I surveys. In the United Kingdom, Phase I and II surveys are not seen as separate stages, and site evaluation is much more the focus. Reconnaisance survey is however carried out in the form of field walking in many areas, and Champion *et al.* (1995) suggest that a survey could combine strategies, by doing a test pit survey first to detect archaeological sites, followed by machine trenching for site investigation. It is important to realize that the goals of reconnaissance and evaluation surveys are different. Reconnaissance survey in the Netherlands is trying to find potential site locations, using indicators such as soil colorations, palaeosols or charcoal to decide whether a site may be near. It is not absolutely necessary to determine the existence and position of the site itself; this may well be done during evaluation. Therefore, reconnaissance survey can focus on indicators that are relatively easy to observe, and can be carried out by means of small unit sampling. However, in order to obtain absolute certainty on the location and size of an archaeological site, machine trenching is the only technique that will provide this certainty.

6.8. CONCLUSIONS

A simple scheme can be made of the relation between the results of a survey and the actual archaeology present in the surveyed area (see also Orton, 2000b:119):

	site observed	**no site observed**
site present	survey successful, complete certainty	survey not successful
no site present	impossible situation	survey successful, but no complete certainty

Too often, results of surveys are presented as if they describe the characteristics of the total archaeological record in an area, whereas in reality we are only talking about a sample of a target population of unknown size. The success rate of a survey is however not determined by the number of archaeological sites discovered, but by the number that has not been discovered; or better said by its failure to determine if a non-site observation is an indication of site absence.

The main factors involved in failure of a survey to discover archaeological sites are the following:

- The sampling units chosen are too small in order to detect the archaeological remains present. It is a factor that is very difficult to control before starting a survey, as it implies that the archaeologist must know beforehand the characteristics of the sites that are looked for. Even though some general

[13] in the Netherlands both types of survey tend to become more and more integrated and are now referred to in the KNA as *inventory survey* (College voor de Archeologische Kwaliteit, 2001)

guidelines can be given in this respect, empirical data on minimal artefact densities that can be observed with different observation methods and sample unit sizes are usually not available.

- Too few observations are made in an area, as a consequence of which the sites that fall between the sampling points (or lines in the case of field survey) will escape detection. Again, it is not easy to control for this factor beforehand, as some prior information on the size of the sites that are looked for should be available in order to obtain the optimal sampling configuration.
- The survey method chosen has not penetrated deeply enough. Establishing the necessary penetration depth of the survey implies that prior knowledge about the local geo(morph)ology is available.

Prior knowledge is therefore necessary to perform a successful survey - a statement that is less paradoxical than it seems. Survey strategies are, economic considerations left aside, often dictated by implicit assumptions about the archaeology and geo(morph)ology of an area; why not make these assumptions explicit before starting a survey? The recommendations put forward by Tol *et al.* (in prep.), and earlier by Zeidler (1995), suggest that a survey should start by defining what type of site one is looking for, and choose for a particular survey strategy depending on the desired accuracy. In this way it becomes much clearer what risks may be associated with a particular survey strategy. It is however still very difficult to relate survey bias control methods to the real world of archaeological sites, as the main factors involved are site size and the density distribution of artefacts and features in sites. These are properties that are in many cases insufficiently known, both in a statistical as well as in a geographical sense. Tol *et al.* (in prep.) attempted to capture these properties for various site types, which resulted in a basic classification scheme of size classes and artefact density classes, some of which had to be based on very scant data. As useful as such a basic classification is, there still is a long way to go before well-established methods of bias control (and correction) will be developed that can be applied to specific site types, and that may also be used to interpret survey data for predictive modelling.

This paper is not intended to offer final solutions for the problem of finding the most effective survey strategy, as there is no such solution. From the sources consulted it can be concluded that statistical theory has had little effect on the practice of archaeological survey design, and standard strategies are prevailing, even if these standards differ from place to place. However, when the theoretical basis of sampling is taken seriously, one cannot confine oneself to standard strategies, and each survey design will have to be based on an analysis of the risks involved with choosing for a particular strategy. This means negotiating between the available resources and the desired accuracy of the survey, a process in which economic considerations will often prevail. A sound statistical foundation of the accuracy of various sampling strategies may then be very helpful in the decision making process. Furthermore, from a political perspective, it may be very important to define explicitly the available prior knowledge that determines the type of survey chosen. It is inevitable that surveys will run into surprises that may be exciting to archaeologists, but are not usually pleasing the contractors. If the survey is based on a specific predictive model, the archaeologist cannot be blamed for finding something that was not expected beforehand. All this means that archaeological heritage managers should not confine themselves to convenient standard survey strategies, but should be able to define the goals of survey in terms of the desired accuracy, keeping a specific hypothesis in mind. It is only in this way that a transparent and balanced decision can be made on where to spend money for archaeological research.

ACKNOWLEDGEMENTS

The author would like to acknowledge that part of this paper is building on the results of the project 'Pilotstudie Boorstrategieën'[14], carried out by RAAP and commissioned by the Senter Agency of the Dutch Ministry of Economic Affaires within the programme Technology and Society. The author would also like to thank his colleague Adrie Tol, and the director of RAAP, Marten Verbruggen, for their valuable contributions. Thanks also go to Thomas Whitley (Brockington and Associates, Inc.) and Hal Darwood (Worcestershire County Council Historic Environment and Archaeology Service) for providing useful information and references concerning survey practice in the United States and United Kingdom. The paper was also improved by the comments made by its reviewer, dr. Mike Baxter (University of Nottingham).

EQUATIONS

1. The probability that a site will be intersected by a transect is given by (Davis, 1986; Sundstrom, 1993):

$$I = \frac{P}{\pi \cdot d}$$

where

P = the perimeter of the site; and
d = the distance between the transects.

2. The probability that a site will be intersected by the points of a regular point sampling grid is given by (Drew, 1979):

$$I = \frac{A}{i \cdot s}$$

where

A = the area of the site;
i = the distance between sampling points; and
s = the distance between sampling rows.

3. The number of samples needed to obtain an estimate of a proportion with a certain level of probability is (Shennan, 1988):

$$n = \frac{Z^2 \cdot p \cdot (1-p)}{d^2}$$

where

Z = the Z-score (number of standard deviations) associated with the desired level of probability;
p = the estimated proportion; and
d = the desired confidence interval width;

[14] 'Pilot study core sampling strategies'

4. The number of samples needed to intersect, with chosen probability, the proportion θ of an area occupied by archaeological remains is given by (Orton, 2000b):

$$n = \frac{log(1-p)}{log(1-\theta_p)}$$

where

p = the desired level of probability; and
θ_p = the upper confidence limit of θ.

5. The probability of finding an artefact in a small unit sample is given by (Stone, 1981; Tol *et al.*, in prep.):

$$D = 1 - e^{-A \cdot d \cdot W}$$

where

e = the base of natural logarithms (= 2.711828);
A = the sample unit's area;
d = the density of artefacts per area unit; and
W = the probability that the artefacts will actually be observed.

Tol *et al.* (in prep.) conclude that the term W should be included in the equation, as it directly influences the proportion of observable artefacts.

BIBLIOGRAPHY

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. Archaeopress, Oxford.

Champion, T., S.J. Shennan and P. Cuming, 1995. *Planning for the past, volume 3. Decision-making and field methods in archaeological evaluations.* University of Southampton / English Heritage, Southampton.

College voor de Archeologische Kwaliteit, 2001. *Kwaliteitsnorm Nederlandse Archeologie, versie 2.0.* http://www.cvak.org

Davis, J.C., 1986. *Statistics and Data Analysis in Geology, Second Edition.* John Wiley and Sons, New York.

Drew, L.J., 1979. 'Pattern Drilling Exploration: Optimum Pattern Types and Hole Spacings When Searching for Elliptical Shaped Targets'. *Mathematical Geology* 11:223-254.

Florida Department of Transportation, 2001. *Cultural Resource Management Handbook.* Florida Department of Transportation, Environmental Management Office, Tallahassee. http://www11.myflorida.com/emo/pubs/cultmgmt/cultmgmt.htm

Georgia Council of Professional Archaeologists, 2001. *Georgia Standards and Guidelines for Archaeological Surveys.* Georgia Council of Professional Archaeologists. http://www.georgia-archaeology.org/GCPA/GCP_S&G_Final.htm

Gilbert, R.O., J.R. Davidson Jr., J.E. Wilson & B.A. Pulsipher, 2001. *Visual Sample Plan (VSP) Models and Code Verification*. Pacific Northwest National Laboratory, Richland (WA).

Groenewoudt, B.J., 1994. *Prospectie, waardering en selectie van archeologische vindplaatsen: een beleidsgerichte verkenning van middelen en mogelijkheden.* Nederlandse Archeologische Rapporten 17. Rijksdienst voor het Oudheidkundig Bodemonderzoek, Amersfoort.

Hey, G. and M. Lacey, 2001. *Evaluation of Archaeological Decision-making Processes and Sampling Strategies.* Kent County Council / Oxford Archaeological Unit, Oxford.

Kintigh, K.W., 1988. 'The effectiveness of subsurface testing: a simulation approach'. *American Antiquity* 53: 686-707.

Krakker, J.J., M.J. Shott and P.D. Welch, 1983. 'Design and evaluation of shovel-test sampling in regional archaeological survey'. *Journal of Field Archaeology* 10: 469-480.

Louisiana Division of Archaeology, 1999. *Investigation and Report Standards.* Louisiana Division of Archaeology, Baton Rouge. http://www.crt.state.la.us/crt/ocd/arch/review/report.htm

McManamon, 1984. 'Discovering sites unseen', in: Schiffer, M.B. (ed.), *Advances in Archaeological Method and Theory 7*. Academic Press, New York, pp. 223-292.

Nance, J.D., 1981. 'Statistical fact and archaeological faith: two models in small site sampling.' *Journal of Field Archaeology* 8: 151-165.

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

Orton, C., 2000a. 'A Bayesian approach to a problem of archaeological site evaluation', in: Lockyear, K., T.J.T. Sly and V. Mihailescu-Birliba (eds.), *CAA96. Computer Applications and Quantitative Methods in Archaeology*. BAR International Series 845. Archaeopress, Oxford.

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

Shennan, S., 1988. *Quantifying Archaeology*. Edinburgh University Press, Edinburgh.

Shott, M.J., 1989. 'Shovel-test sampling in archaeological survey: comments on Nance and Ball, and Lightfoot'. *American Antiquity* 54: 396-404.

Sims, D.C., 2001. *Guidelines for Archaeological Investigations and Reports in Mississippi. Revised Version.* Mississippi Department of Archives and History, Jackson. http://www.mdah.state.ms.us/hpres/archguidelines.pdf

Stone, G.D., 1981. 'On artifact density and shovel probes'. *Current Anthropology* 22: 182-183.

Sundstrom, L., 1993. 'A simple mathematical procedure for estimating the adequacy of site survey strategies'. *Journal of Field Archaeology* 20:91-96.

Tainter, J.A., 1983. 'Settlement Behavior and the Archaeological Record: Concepts for the Definition of "Archaeological Site"'. *Contract Abstracts and CRM Archaeology* 3: 130-133.

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

Verhagen, P. and A. Tol, 2004. 'Establishing optimal core sampling strategies: theory, simulation and practical implications', 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. Archaeopress, Oxford, pp. 416-419.

Virginia Department of Historic Resources, 2001. *Guidelines for Conducting Cultural Resource Survey in Virginia. Revision 2001*. Virginia Department of Historic Resources, Richmond. http://state.vipnet.org/dhr/pdf_files/SurveyManual.PDF

Voorrips, A., S.H. Loving and H. Kamermans, 1991. *The Agro Pontino Project.* Studies in Prae- en Protohistorie 6. Instituut voor Pre- en Protohistorie, Universiteit van Amsterdam, Amsterdam.

Wilson, R.D., 2000. *Air Photo Interpretation for Archaeologists. 2000 Edition.* Tempus Publishing Ltd, Stroud.

Zeidler, J.A., 1995. *Archaeological Inventory Survey Standards and Cost-estimation Guidelines for the Department of Defense.* USACERL Special Report 96/40. US Army Corps of Engineers, Construction Engineering Research Laboratory, Champaign. https://www.denix.osd.mil/denix/Public/ES-Programs/Conservation/Legacy/AISS/usacerl1.html

POSTSCRIPT TO CHAPTERS 5 AND 6

Chapter 5 was written as a prelude to the study on core sampling strategies that was finished and published in Dutch in 2004 (Tol *et al.*, 2004). It was presented at the CAA 2003 conference in Vienna at a time when the project was under way, and therefore only touches the basic statistical issues of core sampling. Chapter 6 was written after completion of the core sampling study, and tries to focus on the implications of using statistical theory for both sampling as well as for predictive modelling. Even though several other authors already have published accounts on optimal sampling procedures, I had to conclude that this has not had much effect on archaeological prospection practice. This may partly be attributed to the fact that standard sampling literature is more concerned with probabilistic sampling than with purposive sampling (see also chapter 7). The difference between the two only became fully clear to me after reading Banning (2002), a source that I did not track down while writing chapters 5 and 6. Predictive modelling depends on probabilistic sampling; archaeological heritage management on purposive sampling. For setting up a good prospection plan for AHM, the predictions of interest are not those concerned with relative densities of sites, but those dealing with the expected dimensions and prospection characteristics of sites. This means that traditional predictive models should only play an important role in deciding whether or not to do prospection, but not in deciding what kind of prospection should be done.

An often heard critique of systematic sampling in archaeology is that it can not detect the unique or unexpected. This is only partly true: it is of course impossible to design a prospection strategy for a site type that has unknown characteristics. However, only in cases where the 'unique' refers to a site that is too small for the sampling grid applied, or is too difficult to detect with the prospection method chosen, it may not be found. Even the smallest and most unobtrusive sites will be detected when we strip the whole study area and sieve all the soil. However, trying to find the 'unique' in this way is time-consuming, and will only result in success once in a while. Again, this points to the importance of clearly defining the objectives of the survey, and of calculating the risks of not finding what we're looking for.

Even though the core sampling report was well received in Dutch archaeology (it has sold quite a number of copies), its recommendations are not yet fully implemented in Dutch archaeological heritage management practice. This may partly be due to the fact that the report does not offer clear-cut recipes on how to deal with site type X or Y. A best practice guideline was recently added to version 3 of the *Kwaliteitsnorm Nederlandse Archeologie*[15]. This will hopefully embed the procedures recommended into everyday archaeological practice. Furthermore, the concepts discussed in chapters 5 and 6 and in Tol *et al.* (2004) have also found their way into the chapter on prospection of the *Nationale Onderzoeksagenda Archeologie*[16], to be published in 2007.

The report by Tol *et al.* contains one minor omission, which also appears in chapter 5: when establishing the optimal survey strategy, one should use the discovery probability as the primary measure for good prospection. I accepted on face value the assumption made by Krakker *et al.* (1983) that discovery probability equals detection probability multiplied with intersection probability. However, the characteristics of the binomial distribution imply that detection probabilities below 1 can never fully guarantee that a site will be found, no matter how many samples are taken. The correct approach to setting up a core sampling strategy is therefore to calculate the number of samples needed to reach an acceptable detection probability, and use this figure to determine the distance between sampling points, dependent on the size of the expected site type. This means that in practice the discovery probabilities of sites are lower than those mentioned in the report.

ADDITIONAL REFERENCES

Banning, E.B., 2002. *Archaeological Survey.* Manuals in Archaeological Method, Theory and Technique 1. Kluwer Academic / Plenum Publishers, New York.

[15] 'Leidraad IVO Karterend Booronderzoek' of the handbook of Dutch Quality Norms for Archaeology (www.sikb.nl/upload/documents/leidraad IVO karterend booronderzoek def.pfd)

[16] National Archaeological Research Agenda

CHAPTER 7 Predictive models put to the test

7.1. INTRODUCTION

7.1.1 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[1]. 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 7.2, 7.3 and 7.4). 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 7.5). In section 7.6, 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 7.7, the conclusions and recommendations of the study will be presented.

7.1.2 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

[1] for an account on what archaeological predictive modelling is about, the reader is referred chapter 1

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 7.4).

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.

7.1.3 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[2] 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[3].

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

[2] 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

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

fossil tidal creek beds, which constitute the highest parts in the landscape[4] (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 to the pipeline from archaeological research carried out during the restructuring program.

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)[5], 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

[4] 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

[5] this is attributed to the poor quality of the pottery, which will weather easily

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 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 defensible 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.

7.2. MODEL PERFORMANCE ASSESSMENT

In this section, a number of methods will be discussed that deal with obtaining a measure of predictive model performance. Sections 7.2.1 to 7.2.4 and 7.2.6 discuss measures and techniques that have been used for model performance assessment in predictive modelling, with varying amounts of success. Sections 7.2.5 and 7.2.7 use some of these methods to judge the performance of some Dutch predictive models. In section 7.2.8 the effects of spatial autocorrelation and spatial association on model performance will be shortly investigated. In section 7.2.9 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') which 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 7.2.4). 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[6] 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: are most of the sites captured by the model? Precision in predictive modelling refers to the ability of the model to limit the area of high probability as narrowly as possible (figure 7.1). 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'.

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 be surveyed. 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 therefore can not be calculated. Gross

[6] referred to as 'specificity' by van Leusen *et al.*, 2005:33

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.

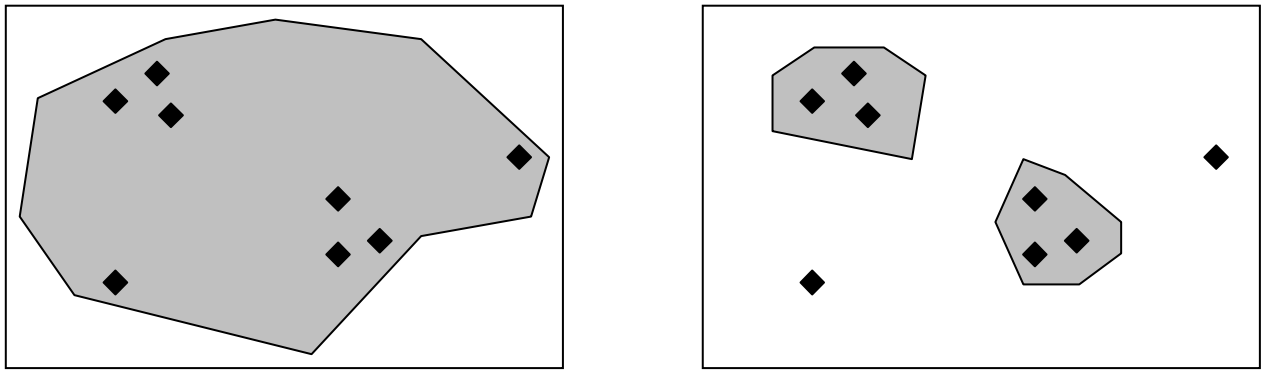


Figure 7.1. The difference between accuracy and precision. The model to the left is 100% accurate: it captures all sites (the black lozenges) in the model (depicted in grey). The model to the right is less accurate, but more precise.

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.

7.2.1 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

p_a= the area proportion of the zone of interest (usually the zone of high probability); and
p_s= 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-measure:

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

where

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[7], 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 7.2.4).

7.2.2 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[8], 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).

---

[7] obviously, the size of the raster cells will influence the value of p_w

[8] the term *classifier* refers to the rule used for obtaining the classification

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 which 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 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 7.2.4 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 7.1-4; 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 7.1. 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 7.2). 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 it is more important to know the probability of finding a site or non-site in each zone. 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 7.3). 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	$P_{m\|s} = 0.85$	$P_{m'\|s} = 0.15$	$P_{m\|s} + P_{m'\|s} = 1.00$
S'	$P_{m\|s'} = 0.20$	$P_{m'\|s'} = 0.80$	$P_{m\|s'} + P_{m'\|s'} = 1.00$

Table 7.2. The conditional probabilities obtained from table 7.1.

	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 7.3. The reverse conditional probabilities obtained from table 7.1.

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 7.1.

	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 7.4. Performance statistics for the model from table 7.1.

$P_{s|m}/P_{s,}$, 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.

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 7.2.8). 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.

7.2.3 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 creating 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.

7.2.4 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 7.5):

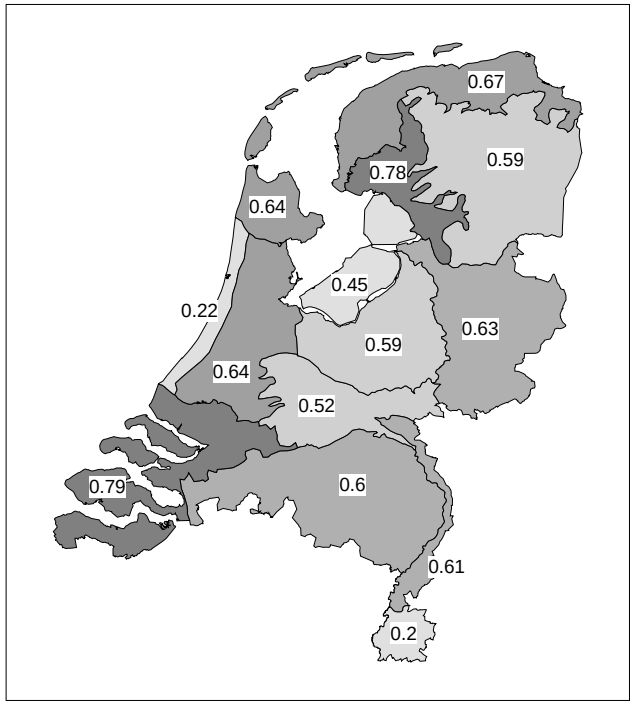


Figure 7.2. Kvamme's gain values per archaeo-region for the 2nd generation IKAW.

	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 7.5. 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 7.2). 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 7.6):

	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 7.6. 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.

The predictive map for the municipality of Ede, province of Gelderland (Heunks, 2001) shows quite a different picture (table 7.7).

	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 7.7. Performance statistics for the municipal map of Ede (Heunks, 2001)

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.

7.2.5 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 were 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)[9]:

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

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):

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

where

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

[9] calculation of the kappa coefficient, but not its variance, is included as a standard tool in Idrisi

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

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

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

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.

7.2.6 COMPARING CLASSIFICATIONS: AN EXAMPLE FROM PRACTICE

In 2003, RAAP Archeologisch Adviesbureau produced a revised version of the IKAW for the province of Gelderland[10]. 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 7.8; figure 7.3).

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 7.9 (the totals are slightly different because of the occurrence of no data zones, that are also defined differently).

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

[10] as part of the map of cultural historical values of Gelderland (Gelderse Cultuurhistorische Waardenkaart, or GCHW)

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.

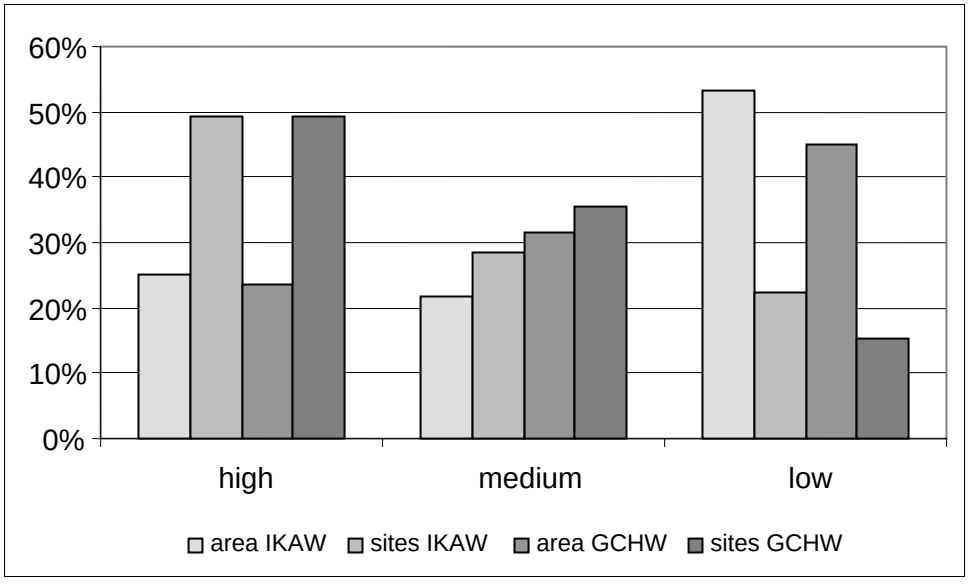


Figure 7.3. Comparison of the IKAW and GCHW maps, based on the screened archaeological data set.

	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, no. sites	3157	1831	1430	6418
%	49.19%	28.53%	22.28%	
GCHW, no. 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. observations	4458	2517	1914	8889
%	50.15%	28.32%	21.53%	
GCHW, no. 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 7.8. Comparison of performance of the IKAW and the GCHW-maps. The comparison is made using both the screened archaeological data-set (#sites) and the original ARCHIS observations.

km²	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%

no. 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 7.9. Comparison of IKAW- and GCHW-classifications.

7.2.7 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[11] and Geary's c[12]. 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 formal publications.

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)[13]. 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.

7.2.8 SUMMARY AND DISCUSSION

In the current section I have so far reviewed criteria and measures for the assessment of model performance (7.2.2 and 7.2.3), and measures and methods for optimising model performance (7.2.4 and 7.2.6), but have refrained from discussing their utility for the task in hand – namely, to give an adequate description of predictive 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

[11] included as a standard tool in Idrisi and ARC/INFO GRID
[12] included as a standard tool in ARC/INFO GRID
[13] incorporated in ARC/INFO GRID as a standard tool

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 7.4; 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[14]. 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 7.10). 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[15]; 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

[14] 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).

[15] 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 zone

planning purposes, as its low 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 7.2.5) 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 7.10. 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.

7.3. 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 7.3.1. However, the procedures can equally well be used with independent data sets. In section 7.3.2 the utility of validation methods for assessing predictive model performance will be discussed.

7.3.1 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 are left out from model building, and methods that re-use parts of the complete data set in order to obtain a more accurate model. 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 were 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 7.2.3.

The simplest resampling method available is *cross validation*[16]. 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

[16] 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

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 7.11 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.

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 7.11. 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 re-use 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 7.11: in fact, resampling can be carried out with subsets of any size. Resampling is also closely related to Monte Carlo-simulation[17], with the difference that Monte Carlo-simulation does not use sample data, but creates 'virtual' samples instead. Table 7.12 summarizes the main differences between the various methods.

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 sampling limit the application of standard statistical inference techniques. Resampling will in those cases generally offer better estimates of the population characteristics than classical inference methods, which rely heavily on the assumption of idealized statistical distributions. 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.

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 7.12. Resampling techniques available for statistical inference.

7.3.2 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.

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

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 are 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 7.3.2 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 7.2.9). 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[18], 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[19] (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.

7.4. 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 become available, we will have to use statistical methods to decide whether we trust our model or not. Section 7.4.1 will briefly discuss the utility of statistical testing methods for predictive modelling, and in section 7.4.2 some examples will be given of testing methods that can be applied to Dutch predictive models.

7.4.1 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 distribution of archaeological sites differs from a by-chance model. We are then comparing the

[18] 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)
[19] the annual conference on Computer Applications and Quantitative Methods in Archaeology

site distribution to an uninformative statistical hypothesis[20], 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 checking 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 7.3. 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 a 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 are 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 the 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

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

and precision in predictive models that has led to the proliferation of performance assessment measures as discussed in sections 7.2 and 7.3.

7.4.2 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 don't 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 7.13). 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 follows:

	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 7.13. 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 7.14).

	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 7.14. Probability of misclassification (p) of IKAW zones, based on the data in table 7.13.

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 7.15). 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 7.15. Probability of misclassification of qualitative probability zones, based on the data in table 7.13.

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 don't 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 7.6.2).

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.

7.5. 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 7.5.1 and 7.5.2). 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 7.5.3 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 7.5.4 will shortly discuss the issue of testing of the independent parameters of the model (the 'environmental data'), after which section 7.5.5 will present the conclusions on the potential of independent testing of predictive models.

7.5.1 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[21]. 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 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 many data are 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 7.4.2, 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 7.5.2 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:

---

[21] 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.

"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 don't know the proportion of the area sampled, which makes it impossible to specify statistical estimates and corresponding confidence limits of site density. In section 7.4.2 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 don't 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.

7.5.2 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.

7.5.3 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. The breakdown of these is as follows:

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 7.16. Breakdown of archaeological observations in ARCHIS according to discovery.

The observations made during archaeological fieldwork can be subdivided into the following categories:

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

So, 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 7.18). 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 7.18 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 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 IKAW-zones or archaeo-regions. The existing data point to a preference of all types of archaeological fieldwork towards the high probability 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 to expectation. However, given the misgivings regarding the quality of the data a quantitative analysis of the research database has not been pursued.

no. 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 7.18. 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 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.

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[22]. 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.

[22] excavations: 18.9% instead of < 7.3%; trial trenches: 21.4% instead of < 9.0%

no. 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
archeological 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 7.19. Number of archaeological observations made in research projects since 1997, subdivided by research type and archaeological probability zone on the IKAW.

7.5.4 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. 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.

7.5.5 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 non-site areas; 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.

7.6. 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 7.6.1 will therefore summarize the appropriate testing methods for the main types of predictive models that are currently in use. Section 7.6.2 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.

7.6.1 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
    - to 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 used to compare different modelling techniques, as has been done by Ejstrud (2003). Obviously, the kappa coefficient (see section 7.2.6) 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.

7.6.2 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"' (translated from Lauwerier & 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 the parcel survey problem described in section 7.5, we don't have predefined units from which to sample[23], 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 7.5; 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

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

[23] 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

data from the Midden-Delfland project[24]. 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 have been collected by Past2Present/Archeologic[25], 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%.

7.7. CONCLUSIONS AND RECOMMENDATIONS

7.7.1 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, 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

[24] kindly put at my disposal by Heleen van Londen (AAC Projectenbureau, University of Amsterdam)
[25] an archaeological consultancy firm, based in Woerden; these data are used here with their permission

  - 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 they become 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 don't 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 are 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. These data have to come from many different sources, and will not all be available in digital form.

7.7.2 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 don't 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 (Rijksdienst voor het Oudheidkundig Bodemonderzoek, Amersfoort)
- Boudewijn Goudswaard (Past2Present/Archeologic, Woerden)
- Dr. Hans Kamermans (Faculty of Archaeology, Leiden University)

- Richard Kroes (Past2Present/Archeologic, Woerden)
- Dr. René Isarin (Past2Present/Archeologic, Woerden)
- Dr. Martijn van Leusen (Institute of Archaeology, University of Groningen)
- 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.

BIBLIOGRAPHY

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

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

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. Archaeopress, Oxford.

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).* Council of Europe, Strasbourg, pp. 181-190.

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. International Institute for Applied System Analysis, Laxenburg. 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. Kluwer Academic / Plenum Publishers, New York.

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

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

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

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.* Center for Archaeological Resource Prediction, Lakehead University, Thunder Bay.

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

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: Willems, W., H. Kars and D. Hallewas (eds.), *Archaeological Heritage Management in the Netherlands. Fifty Years State Service for Archaeological Investigations*. ROB, Amersfoort, pp. 76-118.

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

Delicado, P., 1999. 'Statistics in Archaeology: New Directions', in: Barceló, J.A., 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, Archaeopress, Oxford, pp. 29-38. 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. Rijksdienst voor het Oudheidkundig Bodemonderzoek, Amersfoort, pp. 93-108.

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 Kit*, pp. 33-58.

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.* Taylor and Francis, London, pp. 129-134.

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. Chapman & Hall, New York.

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

Ejstrud, B., 2005. 'Taphomic 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. Rijksdienst voor het Oudheidkundig Bodemonderzoek, Amersfoort, pp. 183-194.

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

Gibbon, G.E., 2002. *A Predictive Model of Precontact Archaeological Site Location for the State of Minnesota. Appendix A: Archaeological Predictive Modelling: An Overview.* Minnesota Department of Transportation, Saint Paul. 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.* Minnesota Department of Transportation, Saint Paul. 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. Rijksdienst voor het Oudheidkundig Bodemonderzoek, Amersfoort, pp. 205-223

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. Geo Books, Norwich.

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

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

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

Hobbs, E. 2003. 'The Minnesota Archaeological Predictive Model', in: Kunow, J. and J. Müller (eds.). *Symposium The Archaeology of Landscapes and Geographic Information Systems. Predictive Maps, Settlement Dynamics and Space*

and Territory in Prehistory. Forschungen zur Archäologie im Land Brandenburg 8. Archäoprognose Brandenburg I. Brandenburgisches Landesamt für Denkmalpflege und Archäologisches Landesmuseum, Wünsdorf, pp. 141-150.

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.* Minnesota Department of Transportation, Saint Paul. 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.* Open Court, Chicago.

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, Rijksdienst voor het Oudheidkundig Bodemonderzoek, Amersfoort, pp. 61-81.

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

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

Kamermans, 2003. 'Predictive Maps and Land Quality Mapping', in: Kunow, J. and J. Müller (eds.), *Symposium The Archaeology of Landscapes and Geographic Information Systems. Predictive Maps, Settlement Dynamics and Space and Territory in Prehistory.* Forschungen zur Archäologie im Land Brandenburg 8. Archäoprognose Brandenburg I. Brandenburgisches Landesamt für Denkmalpflege und Archäologisches Landesmuseum, Wünsdorf, pp. 151-160.

Kamermans, H., J. Deeben, D. Hallewas, P. Zoetbrood, M. van Leusen and P. 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. Rijksdienst voor het Oudheidkundig Bodemonderzoek, Amersfoort, pp. 13-24.

Kamermans, H. and E. Rensink, 1999. 'GIS in Palaeolithic Archaeology. A Case Study from the Southern Netherlands', in: Dingwall, L., 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.

Kohler, T.A., 1988. 'Predictive Modelling: History and Practice', in: Judge, J.W. and L. Sebastian (eds.), *Quantifying the Present and Predicting the Past: Theory, Method and Application of Archaeological Predictive Modelling.* U.S. Department of the Interior, Bureau of Land Management, Denver., pp. 19-59.

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

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

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. Holos-Verlag, Bonn, pp. 257-295.

Kvamme, K. L., 1992. 'A Predictive Site Location Model on the High Plains: An Example with an Independent Test'. *Plains Anthropologist*, 37, 138, pp. 19-40.

Kvamme, K.L., 1993. 'Spatial Statistics and GIS: an integrated approach', in: Andresen, J., T. Madsen and I. Scollar (eds.), *Computing the Past. CAA92 – Computer Applications and Quantitative Methods in Archaeology.* Aarhus University Press, Aarhus, pp. 91-103.

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

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

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

Leusen, M. van, J. Deeben, D. Hallewas, H. Kamermans, P. Verhagen and P. Zoetbrood, 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. Rijksdienst voor het Oudheidkundig Bodemonderzoek, Amersfoort, pp. 25-92.

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

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. Rijksdienst voor het Oudheidkundig Bodemonderzoek, Amersfoort, pp. 169-182.

Moon, H. 1993: *Archaeological Predictive Modelling: An Assessment*. RIC report 106. Resources Inventory Committee, Earth Sciences Task Force, Ministry of Sustainable Resource Management, Province of British Columbia, Victoria. 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*. Sage Publications, Newbury Park.

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

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

Orton, C., 2000b. 'A Bayesian approach to a problem of archaeological site evaluation', in: K. Lockyear, T. Sly & V. Mihailescu-Birliba (eds.), *CAA 96. Computer Applications and Quantitative Methods in Archaeology*. BAR International Series 845. Archaeopress, Oxford, pp. 1-7.

Petrie, J. E. 1998. *The Accuracy of River Bed Sediment Samples*. Virginia Polytechnic Institute and State University, Blacksburg. MSc thesis. 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: Judge, W.J. and L. Sebastian (eds.), *Quantifying the Present and Predicting the Past: Theory, Method, and Application of Archaeological Predictive Modelling.* U.S. Department of the Interior, Bureau of Land Management Service Center, Denver, pp. 173-256.

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. *Post-depositional and research biases. Experiments in the Collection and Analysis of Archaeological Survey Data: The East Hampshire Survey.* Department of Archaeology and Prehistory, Sheffield University, Sheffield.

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

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

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.* www.resample.com/content/teaching/philosophy/index.shtml, accessed on 25-01-2005. Unfinished manuscript.

Smets, P., 1994. 'What is Dempster-Shafer's model?', in: Yager, R.R., J. Kacprzyk and M. Fedrizzi (eds.), *Advances in Dempster-Shafer Theory of Evidence*. Wiley, New York, pp. 5-34. http://iridia.ulb.ac.be/~psmets/WhatIsDS.pdf, accessed 20-10-2005.

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

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

Verhagen, P., 2005. 'Archaeological Prospection 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. Rijksdienst voor het Oudheidkundig Bodemonderzoek, Amersfoort, pp. 109-122.

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

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

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

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

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

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.* Taylor and Francis, London, pp. 5-32.

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

Wheatley, D. and M. Gillings, 2002. *Spatial technology and archaeology: the archeaological applications of GIS.* Taylor and Francis, London.

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. Rijksdienst voor het Oudheidkundig Bodemonderzoek, Amersfoort, pp. 123-137.

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.* Stadtarchäologie Wien, Vienna (CD-ROM).

PART 3 ALTERNATIVE WAYS OF PREDICTIVE MODELLING

In this part, I have grouped three papers that are each looking at different ways to create archaeological predictive models.

In chapter 8 a case study is presented where an attempt is made to reconstruct the potential agricultural production zones in the Vera Basin in south-eastern Spain. The research questions involved were dealing with long-term land degradation, and therefore the modelling has been performed for all relevant archaeological periods, from the Neolithic till the end of the Arab rule. On the basis of estimates of population size of each archaeological site, hypotheses on dietary needs, the crops cultivated, the productivity of the various landscape units in the area, and the accessibility of the area, a reconstruction of the maximum extent of the potential cultivation zone was made for each site, including the effects of irrigation on productivity. The modelling supported the hypothesis of food supply problems during the El Argar-period (Bronze Age). In order to feed the estimated population size, the model indicated that unsuitable areas had to be taken into production, which may have led to deforestation and increasing erosion. For the Roman period, it turned out that the area cannot have generated surplus production without the introduction of irrigation systems. However, it is known from historical sources that the area exported agricultural produce in this period. It was also noted that the area necessary for cultivation in the Arab period was relatively small, due to the use of refined irrigation systems.

While this study did not aim to produce a predictive model like the ones presented in part 1, the approach offers the potential to set up different scenarios, and to compare these to archaeological reality. One notable aspect of the modelling is the low predicted land use in the vicinity of the town of Turre. No archaeological sites are known in this zone, whereas according to the model Turre's surroundings are eminently suited for agriculture. It may well be that Turre is built on earlier settlements, that for this reason have not been discovered through field survey.

In chapter 9 a review is given of the applicability of land evaluation for creating land use models like the one in chapter 8. Land evaluation is a technique that uses measurable characteristics of the landscape, like soil moisture retention capacity and fertility, to estimate the suitability of soils for agriculture. In archaeology this technique has not been applied frequently, although it can in itself be a useful method to predict agricultural potential, and as such can contribute to the prediction of possible archaeological site locations. In practice however, carrying out a prehistoric land evaluation is far from easy. A few examples are given to make clear that the information necessary to carry out land evaluation is difficult to obtain for the past. Changes in hydraulic regime and erosion for example can strongly influence the agricultural potential. These changes will have to be reconstructed in order to be able to perform a prehistoric land evaluation. Furthermore we cannot assume that the criteria used for modern land evaluation are the same as those applied in the past. Roman sources indicate that workability of the soil was a much more important criterion in the past than it is today. This means that a 'Roman' land evaluation, using the same information, will yield a different outcome from a modern one.

Finally, chapter 10 assesses the potential of integrating social and cultural factors in archaeological predictive modelling. While the landscape is a very important factor in predicting archaeological site location, it cannot provide a full explanation of why people chose to settle somewhere. Social and cultural factors, like the proximity of other settlements, or the presence of religious sites, will have played an important role as well. This has always been an important objection to the current practice of predictive modelling. However, very

few studies are known that try to integrate the socio-cultural component in the models. In this chapter the problem is shortly explained, and some directions are given on how to include social and cultural factors, without resorting to vague concepts like the prehistoric perception of the landscape, but by using measurable aspects, like the accessibility of settlements, the visibility of landscape features with a possible symbolic function, and the continuity of occupation. Such an approach is practically possible, and will not only lead to predictive models with a stronger scientific foundation, but also to better predictions.

CHAPTER 8 Modelling Prehistoric Land Use Distribution in the Río Aguas Valley (S.E. Spain)[1]

Philip Verhagen, Sylvia Gili[2], Rafael Micó[2] and Roberto Risch[2]

8.1. INTRODUCTION

The Río Aguas Project is an international, interdisciplinary research project, funded by Directorate General XII of the Commission of the European Union. Its main goal has been the investigation of the long term evolution of human and natural systems in the Lower and Middle Río Aguas Valley, an area covering approximately 16 by 10 km in the Spanish province of Almería (figure 8.1a and 8.1b). This socio-natural evolution has been a crucial factor in the development of land degradation in the area, which was already studied within the larger context of the Vera Basin for the Archaeomedes Project (van der Leeuw, 1994). The Río Aguas Project has considerably expanded this knowledge through climatic, geomorphologic, hydrological, palaeo-ecological and archaeological investigations. One of the main questions to be addressed was the relationship between the potential resources in the area and the use that has been made of these resources during different (pre-)-historical periods. Agricultural land use in particular has exerted a strong pressure on the environment during such diverse eras as the Bronze Age, the Roman period and the nineteenth century. In order to further investigate the environmental impact of agricultural subsistence strategies, a GIS based modelling of the possible production landscape in the past was undertaken. An outline of the procedure followed can be found in figure 8.2.

Figure 8.1a. Location of the study area in Spain.

---

[1] This paper also appeared in L. Dingwall, S. Exon, V. Gaffney, S. Lafflin and M. van Leusen (eds.), 1999: *Archaeology in the Age of the Internet – CAA97. Computer Applications and Quantitative Methods in Archaeology 25th Anniversary Conference, University of Birmingham.* British Archaeological Reports, International Series 750. Archaeopress, Oxford. CD-ROM.

[2] Departament de Prehistòria, Universitat Autònoma de Barcelona, Edifici B, Bellaterra (Barcelona), Spain. The text of this paper was prepared together with my colleagues from Barcelona on the basis of the GIS modelling I carried out. My Spanish colleagues specified the input to calculate the number of hectares needed for each site, and provided all the archaeological data.

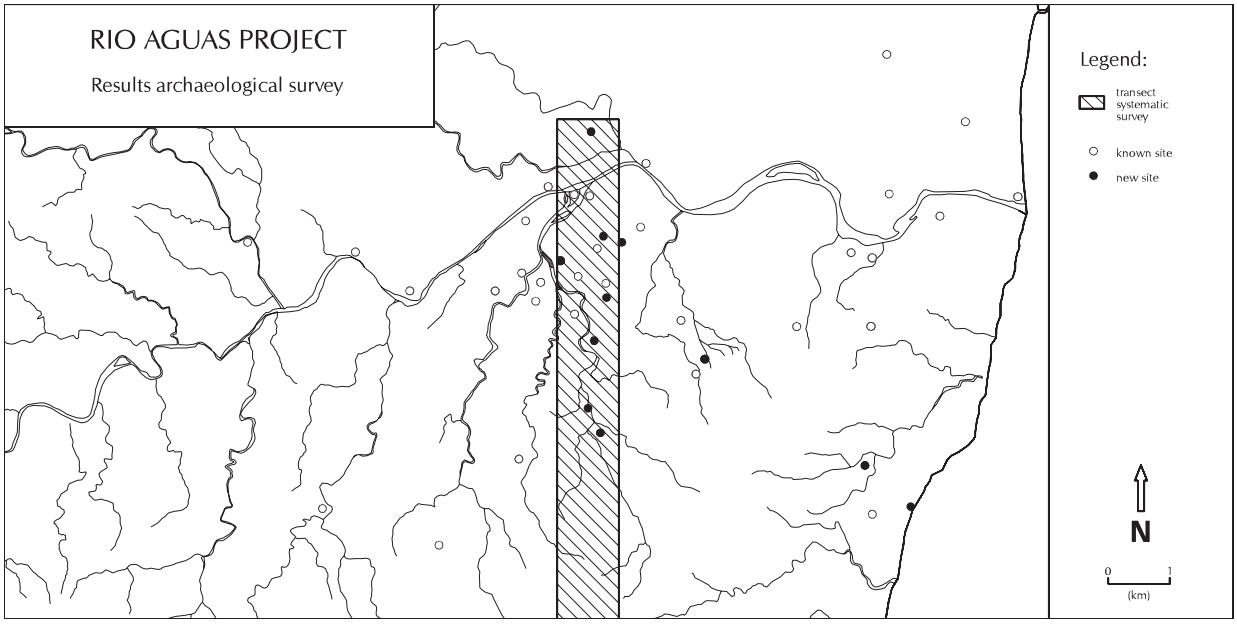


Figure 8.1b. Overview map of the study area.

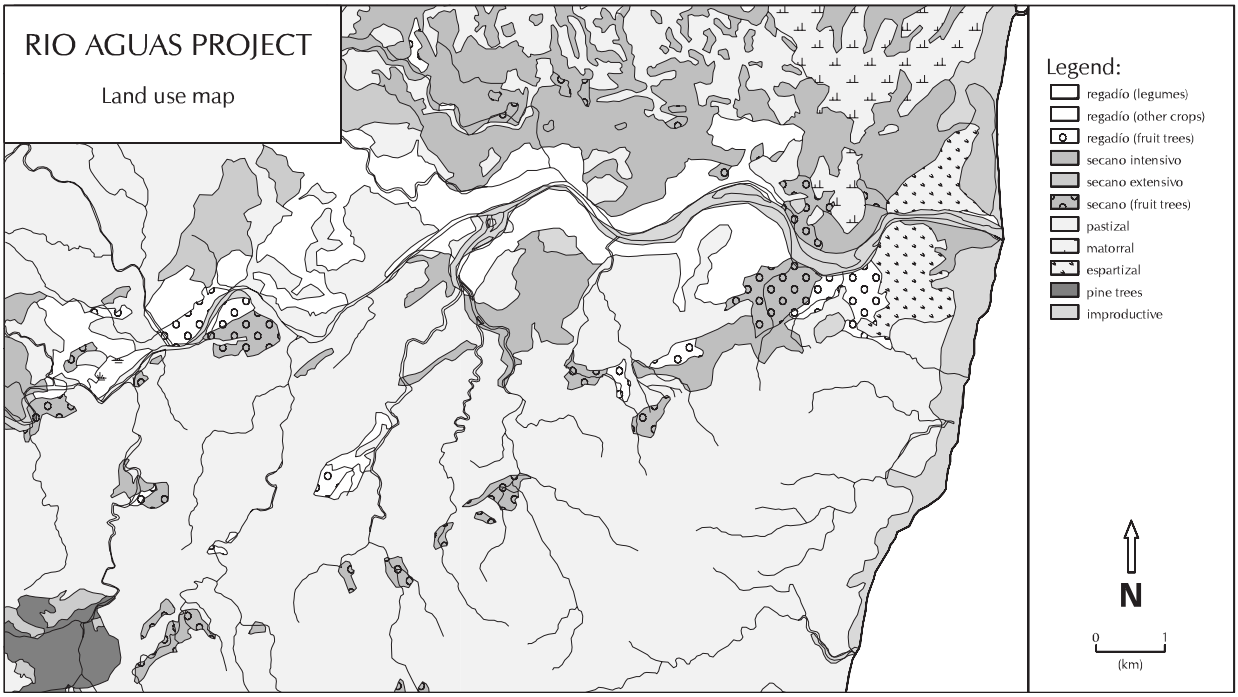


Figure 8.3. Land use map of the area (situation in 1978).

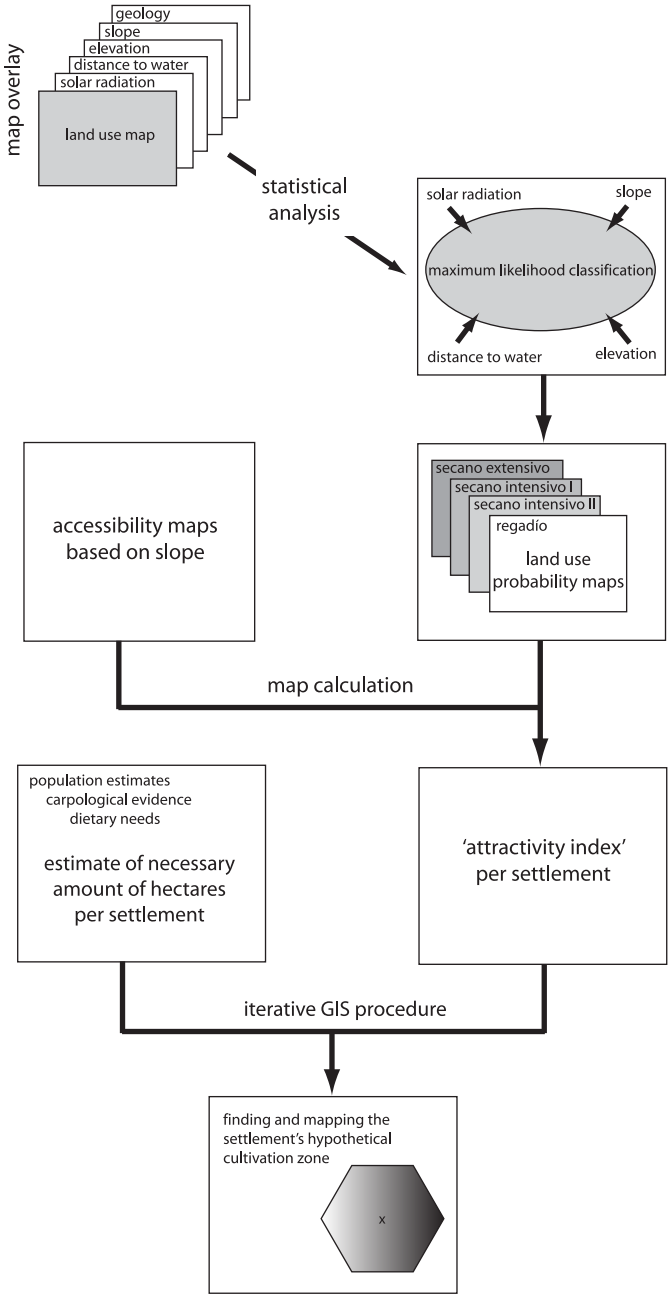


Figure 8.2. Outline of the modelling procedure followed.

8.2. ENVIRONMENTAL CONTEXT

The study area is located in one of the driest regions on the European continent, with rainfalls generally between 250 and 300 mm per year. Annual figures may vary considerably because of the highly irregular nature of the rainfall. The area is dominated by the Sierra Cabrera mountains south of the Río Aguas, which rise steeply from the valley floor to their summit at 934 meters. Geologically speaking, the *sierra* is composed of hard Palaeozoic metamorphic rocks, whereas the foothills consist of softer Miocene marls, sands and limestones. The Río Aguas valley itself is characterised by a broad Quaternary plain with dispersed remnants of Pleistocene river terraces and glacis (Schulte, 1996).

Agricultural land use nowadays is mainly found along the Río Aguas, with irrigated farming of barley, vegetables and fruit trees. On the river terraces irrigation is less frequent, and the main crops here are barley, wheat, almonds and olives. The shallow soil profiles on the northern flanks of the Sierra Cabrera are less used for agriculture, as they are very dry and vulnerable to erosion. Where slope is more gentle, cultivation of almonds, olives and sometimes barley can be found. The ubiquitous terracing found on the intermediate altitudes in the Sierra Cabrera is a remnant of the subsistence practices that existed until very recently in this area and date back as far as the Arab period. The terraces were constructed to catch the water and fine sediment that comes down through the dry river beds during rain storms. These terraces have been used to grow olives, almonds and barley, but most of them have now been abandoned. The central *sierra* area, covered by *garrigue*, was until recently used as grazing land for goats, but this practice has almost come to an end.

8.3. ARCHAEOLOGICAL CONTEXT

The Vera Basin, where the Río Aguas valley is located, presents one of the most complete archaeological records of the western Mediterranean, especially for the later prehistoric periods (Lull, 1983; Chapman, 1991; Castro *et al.*, 1994; Castro *et al*., 1997). Human occupation started in the Neolithic around 5,000/4,500 BC. This first occupation phase was characterized by a subsistence strategy of low intensity and high diversity. Only few settlements have been reported, and they seem to be short-lived. There is no sign of environmental perturbations in this period, either as a consequence of the diversified land management or as a function of the low population density. The ensuing Chalcolithic period (3,000-2,250 BC) shows a steady increase in the number, size and duration of the settlements. Subsistence production is still based on diversified exploitation. The larger and longer-lived settlements are involved in specialised production of metal, stone and bone tools, and exchange networks develop in order to distribute the goods.

Around 2,250 BC a new socio-economic and political system emerges, known as the El Argar culture. The steadily increasing population is concentrated in fewer and permanent settlements, that are located on higher ground. The agricultural production obtained in the plains is concentrated, processed and redistributed in these settlements. However, only a new political class, that also controls the production of metal goods, is benefiting from this system, evidence for which is found in the rich grave goods and higher life expectancies of this group. Towards the Late Argaric period (1,750-1,550 BC) the system develops into what has been referred to as the first state society in the western Mediterranean. Extensive barley cultivation ensures the basic subsistence needs of the population, but only at the cost of deforestation of the lowlands and decreasing health of part of the population. This form of socio-economic organisation, possibly coupled to aridification and geomorphic instability, eventually leads to the sudden collapse of the Argaric society around 1,550 BC.

After the Argaric period, a long lasting process of depopulation starts. Nevertheless, the lowlands do not seem to recover from the deforestation and intensive exploitation suffered during the Argaric period. A new diversified subsistence system develops, introducing the cultivation of olives and grapes, and possibly of small-scale terracing and irrigation infrastructure as well. The establishment of the large settlement of Villaricos in the northeast of the Vera Basin around 800 BC to exploit the lead and silver ores there does not lead to a demographic recovery in the Río Aguas valley, presumably because of the absence of these ores in the Sierra Cabrera.

Only with the integration of the area into the Roman Empire (0-400 AD) does a new phase of economic and social development start. The plains and valleys are exploited through several *villae*, two of which are located along the Río Aguas. Alongside these vast estates, small rural settlements develop, both in the flood plains as well as on the mountainsides, and population increases until the fifth century AD. Agricultural production consists mainly of extensive cultivation of cereals and olives. This period also witnesses the introduction of irrigation systems in order to increase agricultural production in the river valleys. During the Late Roman and Byzantine period (400-650 AD), this situation is largely continued.

After the political and economic crisis of the seventh century AD, a demographic decline and the abandonment of many small settlements is observed, a situation that continues well into the first century of the Arab period (711-1492 AD). The establishment of the Omayyad state leads to a new phase of socio-economic development. Extensive crop production, development of irrigation and terracing systems, mining and iron smelting form the economic basis for a growing population that settles in the lowlands as well as in the mountains close to mining and water resources.

In the thirteenth century, the Vera Basin becomes a border region of the Nazarí state, the last Islamic state in Spain. This may have led to the establishment of new settlements on better defendable sites in the mountains. A sophisticated terracing and water storage systems allows for high agricultural yields, so each settlement can provide for its own subsistence needs. From an environmental point of view, this agricultural system was very efficient for the prevention of further degradation of the area.

8.4. AGRICULTURAL POTENTIAL OF THE RÍO AGUAS VALLEY

One of the aims of the modelling was to come up with a classification of the landscape in terms of agricultural potential which could be used as the basis for land use distribution mapping. The traditional land management system in the area consists of two types of agriculture: *regadío* or irrigated farming, and *secano* or dry farming. Evidently, *regadío* is a more productive land use type than *secano*. However, the area suited for low technology *regadío* is relatively small: it is restricted to the flood plain of the Río Aguas, where irregular flooding provides the high groundwater table necessary for higher yields. This strategy was possibly used since Neolithic times. Two different strategies (that can be used complementarily) have been applied in the past to increase the area suitable for *regadío*: the construction of hydraulic structures like canals and water conduits to transport the water to the fields, and the application of terracing to catch the water before it runs off into the valley. The application of irrigation by means of canals started in the Roman period, whereas the use of terraced agriculture is the trademark of the Arab Nazarí period. When we look at the land use map of the area based on the situation of 1978 (Ministerio de Agricultura, Pesca y Alimentación, 1982; figure 8.3), we can see that the basic subdivision in *regadío* and *secano* is still valid. It is assumed that the location of the regadío and secano areas will depend on the suitability of the terrain for either type of agriculture. By comparing the 1978 land use distribution with the environmental variables available in the Río Aguas GIS (geology,

elevation, slope, solar radiation received, distance to streams) it was possible to arrive at the following scheme of four land use types:

- *regadío* in the Río Aguas flood plain (no artificial irrigation applied) - the basic form of *regadío*, available in all (pre)historical periods
- *secano intensivo I* at larger distances from the river in the Río Aguas flood plain - these areas can be irrigated by means of canals
- *secano intensivo II* on the river terraces of Río Aguas and in the foothills of the Sierra Cabrera - these areas will need a more complex hydraulic infrastructure for irrigation
- *secano extensivo* at intermediate altitudes - a low productivity type of *secano* that has not been very important in prehistory, as the yields are extremely low; this land use type was predominantly applied during the nineteenth century, when demographic pressure forced farmers to take marginal land into production

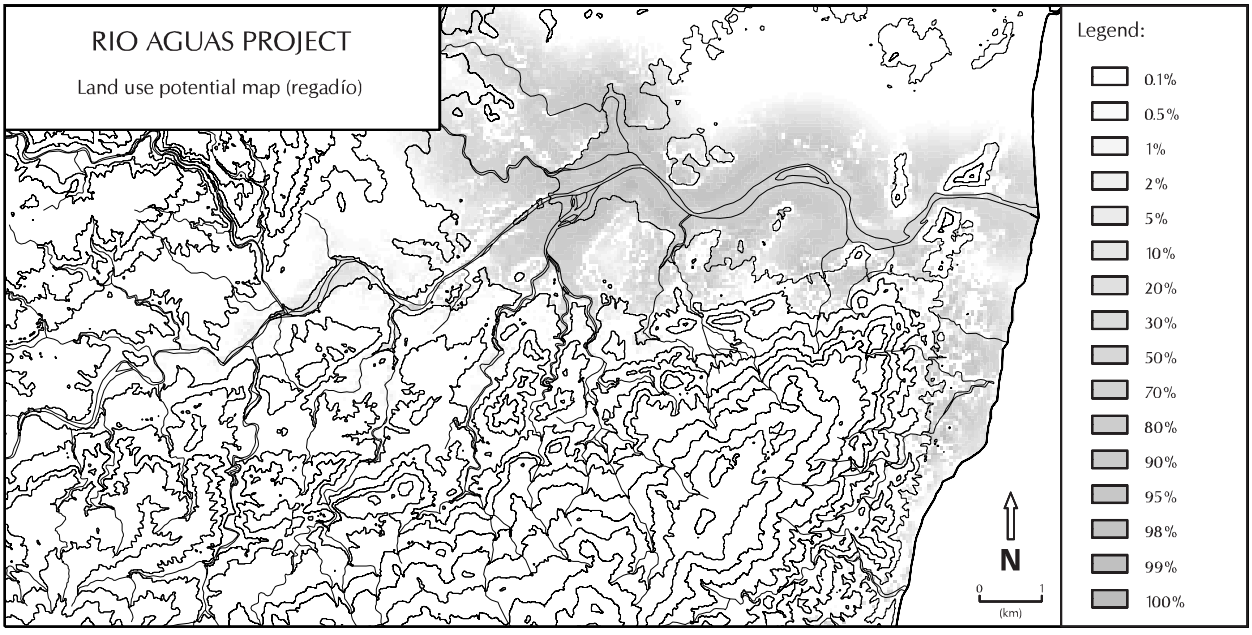


Figure 8.4. Land use probability map for regadío.

In order to find out where these land use types are most probable, a maximum likelihood classification (Schowengerdt, 1983) was performed. Each of the land use types was compared with the available continuous variables (elevation, slope, solar radiation and distance to streams) in order to construct the 'signature files' that describe the characteristics of the land use types with regard to the variables used. The resulting classification yielded probability maps for each land use (see figure 8.4 for an example). By comparing these maps, it can be concluded which land use is the most probable at a certain location, and whether it is probable at all. The results at the 80% probability limit show that approximately 3,000 hectares of agricultural land are available (900 ha *regadío*, 750 ha *secano intensivo I*, 500 ha *secano intensivo II* and 750 ha *secano extensivo*). The resulting map is of course only a tentative approximation of the actual land use potential, but is nevertheless useful as a basic model to define the relation between potential land use and exploitation in

different periods. A surprising conclusion is that there is still a considerable amount of relatively high potential *regadío* and *secano intensivo I* land that is currently not being used.

8.5. LAND SUITABILITY: A FUNCTION OF POTENTIAL AND ACCESSIBILITY

The land use potential map in itself is not sufficient to interpret the attractivity of the land for agriculture. For the modelling, the basic assumption was made that people preferred high potential land at a short distance from the settlement. In order to reconcile these demands, a map was constructed for each single settlement that combined the land use probability map with an accessibility map. The accessibility map is based on the slope of terrain, as this will have been the major constraint to transport by foot in the area (the river beds are dry most of the year). According to Gorenflo and Gale (1990) the effect of slope on travelling speed by foot can be specified as:

$$v = 6\, e^{-3.5 | s + 0.05 |}$$

where:

v = walking speed in km/h
s = slope of terrain, calculated as vertical change divided by horizontal change, and
e = the base for natural logarithms.

This function is symmetric but slightly offset from a slope of zero, so the estimated velocity will be greatest when walking down a slight incline. As we are interested in the time needed to go from a settlement to the fields and back, we add the estimate for going down and going up to find the total amount of time spent. In this particular case an exact estimate is not required, as we will be comparing the relative accessibility of areas, not their absolute values. Using the equation above, a cost surface has been constructed that specifies the amount of time needed to traverse each grid cell and go back again. This cost surface has been used to calculate a cumulative cost surface for each single settlement in order to arrive at a measure of accessibility of the terrain as perceived from the settlement.

The maps of land use probability and accessibility were then combined in a map of ‘attractivity indices’, which are defined as follows:

$$A = p(L)\,(1 - D/D_{max})^2)$$

where:

A = attractivity index
p(L) = land use type probability, measured on a scale from 0 to 1
D = distance, measured in hours walking, and
D_{max} = maximum possible distance.

In this particular case, the distance decay is assumed to be a square function of the actual distance, an assumption commonly used for gravity models, as the travelling time will become increasingly constraining at

larger distances from the settlement. For ease of calculation, D_{max} is set to two hours[3], which means that areas that require more than four hours walking a day in order to be exploited will not be available for cultivation.

The index will range from values near 1 on locations near the settlement with optimum land use potential, to 0 on locations that are either wholly unsuited for agriculture, or that are too far away from the settlement.

8.6. ESTIMATION OF LAND SURFACE NEEDED FOR AGRICULTURE

Having a map of land attractivity for each settlement, the next question to be addressed is the amount of hectares that each settlement needs for its agricultural production. For the modelling, is was assumed that during all periods the settlements will have tried to adopt a strategy of self-sufficiency. Although this is not true for each period, it serves a clear purpose: if we can model the amount of land needed for self-sufficiency and compare this to the actual land available, it is possible to see if there is potential for surplus production, or inversely, if there is insufficient space to grow crops for the whole population. The real uncertainty we are dealing with is the size of the population. The population estimate applied here is following Renfrew (1972):

$$P = 200\ A$$

where:

P = estimated population of the settlement, and

A = the size of the settlement, measured as the extent of the archaeological remains visible on the surface.

Where additional information was available from excavations, this information was used to adapt the population estimates. It is however acknowledged that this method of population estimation cannot lay a claim to extreme accuracy.

The population estimates were then used for the calculation of the number of hectares that needed to be cultivated in order to feed the population. This was done by taking into account the nutritional value of the five most important crops that have been identified by analysis of the seeds found on the archaeological sites. Two types of cereals (barley and wheat) and three types of legumes (beans, peas and lentils) seem to have been important elements of the diet, be it in changing proportions over time (Clapham *et al.*, 1997). Using the nutritional value of each species and its relative importance in the diet as inferred from carpological analysis, it was possible for each period to calculate the amount of each species needed to feed a person, taking an average nutritional need of 2,600 Kcal per person per day. From this amount it was possible to calculate the number of hectares needed for the whole population, taking into account the following conditions:

- grinding of cereals implies a loss of 30% of the original seed weight;
- a certain amount of seed must be stored in order to have a crop next year; the volume of stored seed is different for cereals and legumes, and is also different for secano and regadío; figures from the literature have been applied to calculate the stored seed volume;

[3] most studies on site catchment analysis adopt a limit of 30 minutes to one hour as a maximum (e.g. Chisholm, 1962). The two hour limit, while also used by Gilman and Thornes (1985), was only chosen for ease of calculation. In the model specified, the attractivity of land further away from the site drops rapidly, so the difference in attractivity between land at a distance of one hour and two hours walking is relatively small. No attempt was made to compare between different radii of walking distance.

- for secano, each year of cultivation is followed by two years of fallow in order to recover soil fertility; this implies a multiplication by three of the land needed for secano;
- productivity indices are different for secano and regadío; productivity indices from present day traditional agriculture have been applied; it is unlikely that these figures overestimate the productivity in the past.

In order to make the final calculation, it was necessary to specify the relative importance of regadío and secano for each time period considered.

- from the Neolithic until the Roman Republican period, the evidence suggests that legumes were cultivated on the more humid Río Aguas flood plain, whereas cereals seem to have been cultivated under dry farming;
- from the Roman Imperial period onwards it seems that legumes and wheat were cultivated using irrigation systems, whereas barley continued to be a dry farming crop;
- from the Nazarí period until modern times historical data provide detailed information on the cultivation practices in the area.

For a more detailed description of the method followed and the empirical evidence that supports it, see Castro *et al*. (1996, 1997).

8.7. FINDING THE LAND

Using the hectare calculations and the maps of attractivity indices it is possible to map the hypothetical cultivation zone of each settlement. In the simplest case, this mapping assumes that each settlement has free access to all the land in the study area, and can therefore choose the best land available for its agricultural production. The model was run for three different cases, the number of options increasing with each run: firstly with just the land use potential map for *regadío*, secondly using the land use potential map for both *regadío* and *secano intensivo I*, and thirdly also including *secano intensivo II*. The *secano extensivo* case has not been considered here, as there is no conclusive evidence that this land use type was applied before the nineteenth century. Generally speaking, large cultivation zones will result in lower mean attractivity indices. Sites that were relying on *regadío* alone will not exhibit large differences in mean attractivity between the three runs, whereas settlements that relied on one of the *secano* strategies will show significantly higher attractivity indices when these options are included. So the comparison of the attractivities will tell us if the settlements were in a favourable position for agriculture, and if they could improve their options by including land that is more suitable for *secano* than for *regadío*.

The final step in the modelling was the cartographic representation of the modelled cultivation zones. In order to present a hypothetical land use distribution map for each period, an additional condition had to be fulfilled: the cultivation zones of the settlements should not overlap with the land of neighbouring sites. Basically, two theoretical models can be applied: one that favours the larger sites over the smaller ones (i.e. a form of social exploitation), and a second model that allows the smaller settlements to have their own land, leaving the less attractive bits to the larger ones. This second approach was chosen for the mapping, as it is assumed that social exploitation was the exception and not the rule.

The model should be able to check for each single settlement whether it is trying to claim land that is also desirable to other settlements. As the mapping is performed using an iterative procedure, it is possible to

check on each iteration if there is a potential land use conflict, and let the smallest settlement 'win'. In practice however, this is not the most straightforward procedure, as it requires some fairly complex programming and a considerable amount of computation. Therefore it was decided to start calculating the cultivation zones for each settlement separately, starting with the smallest one and moving up to the largest settlement. A comparison of the two technical solutions showed very little difference in the configuration of the cultivation zones.

8.8. RESULTS

NEOLITHIC PERIOD (5,000/4,500 – 3,000 BC; figure 8.5)

The small settlement sizes in this period lead to small cultivation zones with relatively high attractivity indices. It is nonetheless obvious from the mapping that some sites are trying to use the same land. The explanation for this is simple: not all sites existed simultaneously; the settlements were short-lived, and the same area could be colonised more than once. Most settlements can find sufficient land in the Río Aguas flood plain, where large areas are still left uncultivated. Dry farming may have been important for settlements that were placed in more remote locations. The placement of the settlement of La Raja de Ortega at a distance of approximately 2.5 km north of the Río Aguas is unlike the location of any other settlement. The land use potential of this area between the Río Aguas and Río Antas is very low because of the absence of surface water. As the palaeo-climatic evidence indicates a more humid phase during the early Holocene, this area may have become less attractive in later periods than it was during the Neolithic.

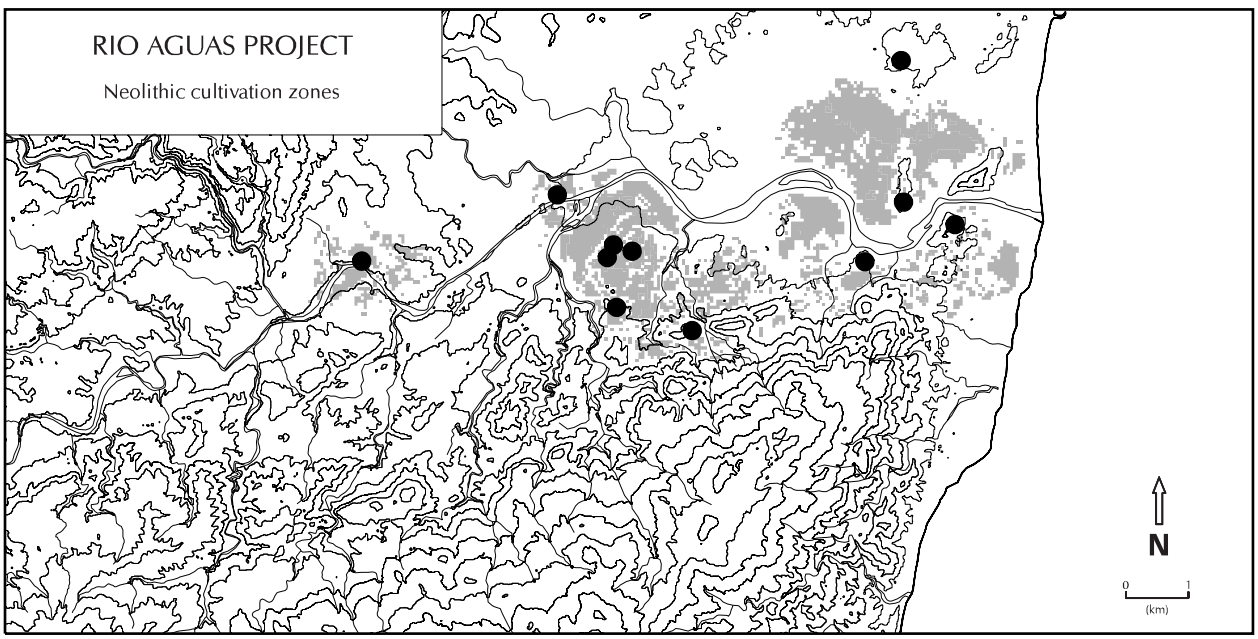


Figure 8.5. Hypothetical cultivation zones for the Neolithic period.

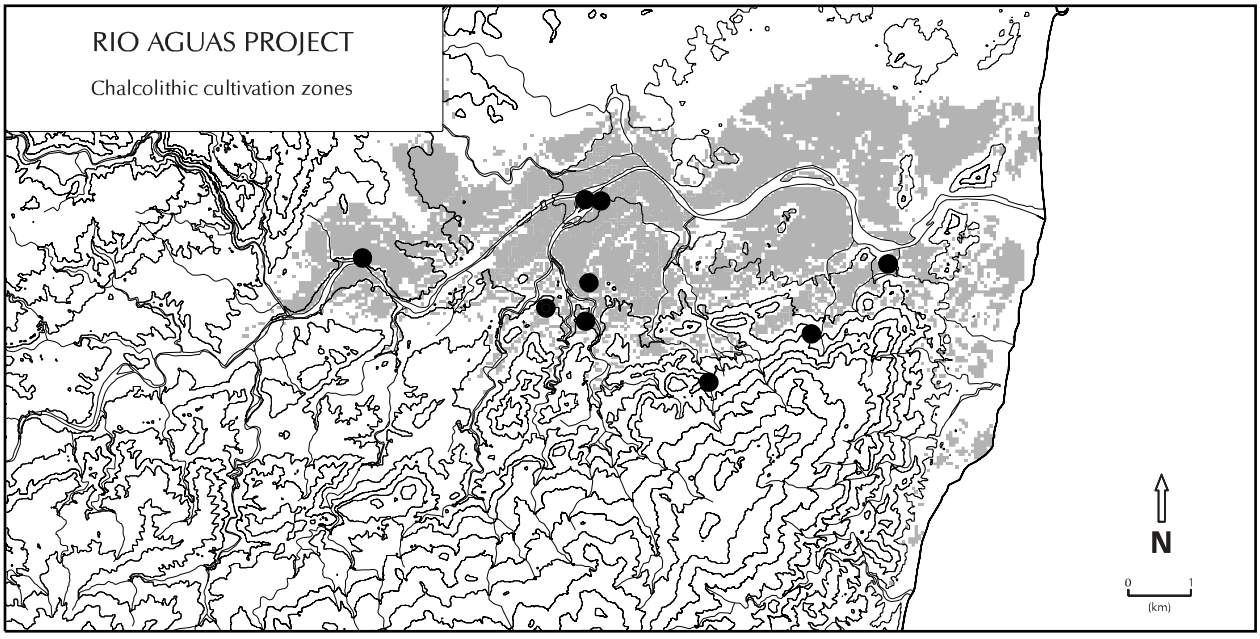


Figure 8.6. Hypothetical cultivation zones for the Chalcolithic period.

CHALCOLITHIC PERIOD (3,000 – 2,250 BC; figure 8.6)

Increasing settlement size leads to larger cultivation zones, which cover almost the entire Río Aguas valley. As in the Neolithic, not all settlements existed at the same time, which may explain the land use distribution pattern around the Rambla de Mofar. The land that is taken into cultivation is located at larger distances from the site than during the Neolithic, and dry farming seems to have become more important.

ARGARIC PERIOD (2,250 – 1,550 BC; figures 8.7, 8.8 and 8.9)

At the transition of the Chalcolithic to the Bronze Age we witness a drastic change: the number of settlements is reduced to four. These are much larger, and are located at greater distances from the Río Aguas flood plain. The archaeological evidence indicates that barley monoculture becomes the dominant cultivation strategy. This implies the need for large cultivation zones for dry farming, which can not be found close to the settlements. This is clear from the mapping for the first phase of the Argaric period (2,300 – 1,900 BC): the coastal zone south of the Río Aguas is taken into cultivation for lack of any better alternative close to the settlements of Barranco de la Ciudad and Peñón del Albar. It is highly improbable that this configuration reflects the actual situation during the Argaric period; a system of exchange will have developed with other settlements in the Vera Basin. During the second Argaric phase (1,900 – 1,750 BC) an attempt is made to increase the production of legumes, which could only take place by means of *regadío*. This decreases the size of the cultivation zones, although they are still located at large distances from the settlements. However, in the last Argaric phase (1,750 – 1,550 BC) a return is observed to barley monocropping. The settlement of Gatas, which has a growing importance during the Argaric, now becomes the centre of barley storage, processing and redistribution. The disastrous consequences of this development are well illustrated by the cultivation zone

mapping for this period: even with the coastal zone in use for agricultural production, almost the complete Río Aguas valley is used for cultivation, including unattractive areas that were not considered in the Neolithic and Chalcolithic. It is highly probable that food production could not keep up with the population growth, which eventually led to the collapse of the Argaric socio-economic system.

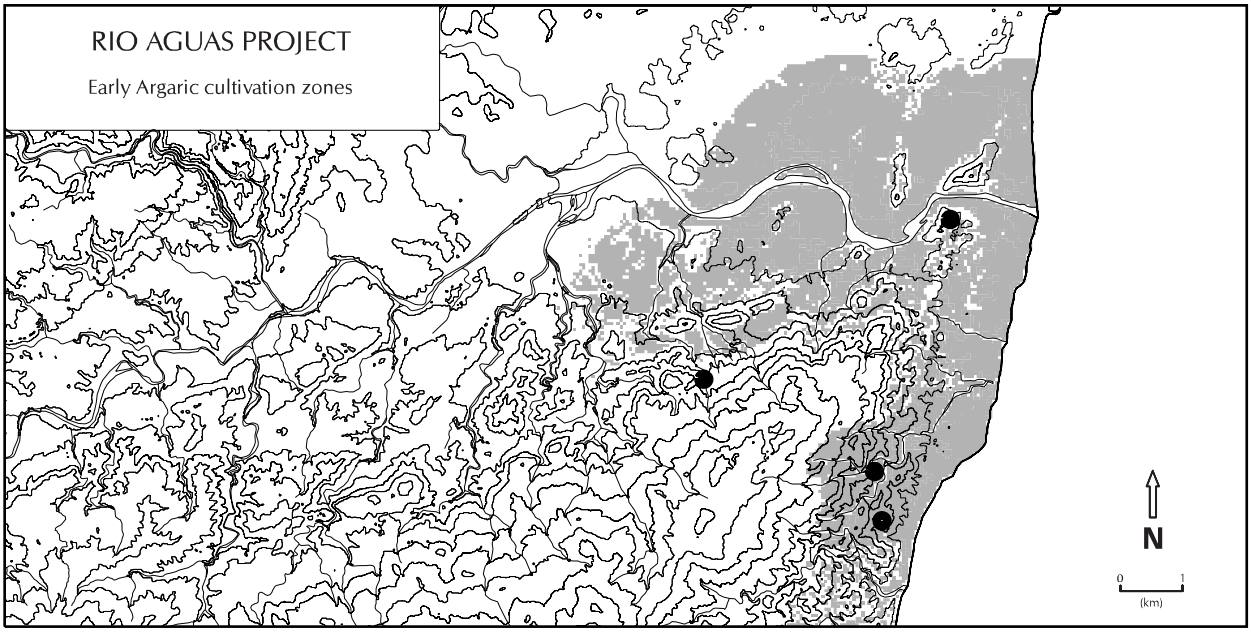


Figure 8.7. Hypothetical cultivation zones for the Early Argaric period.

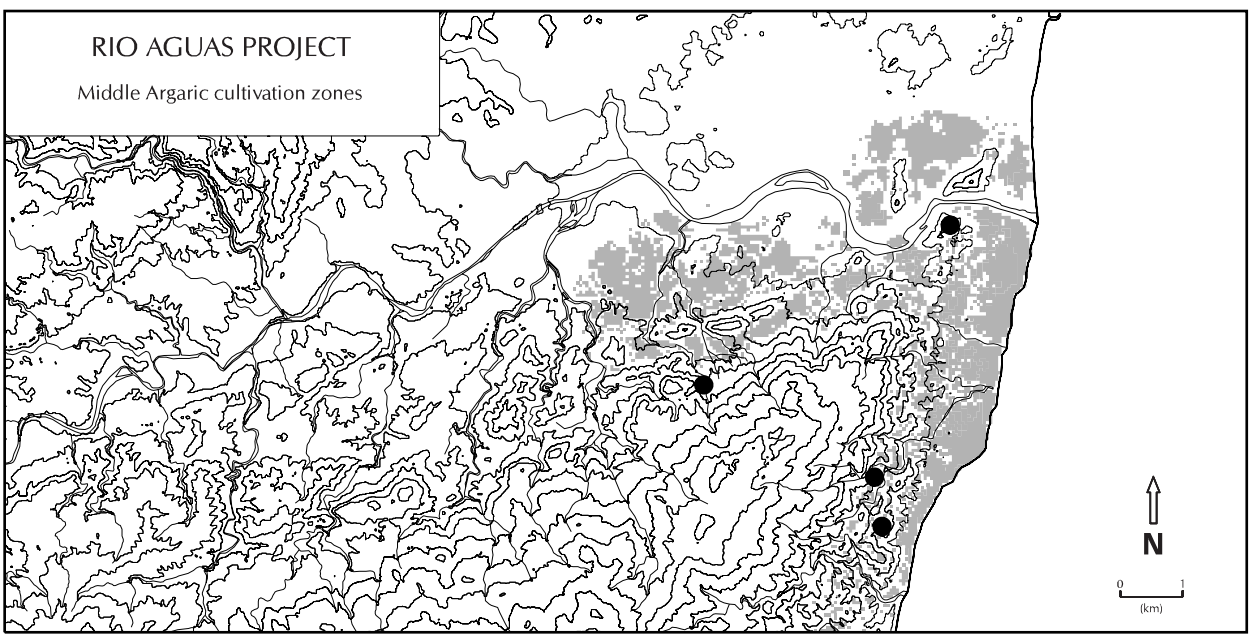


Figure 8.8. Hypothetical cultivation zones for the Middle Argaric period.

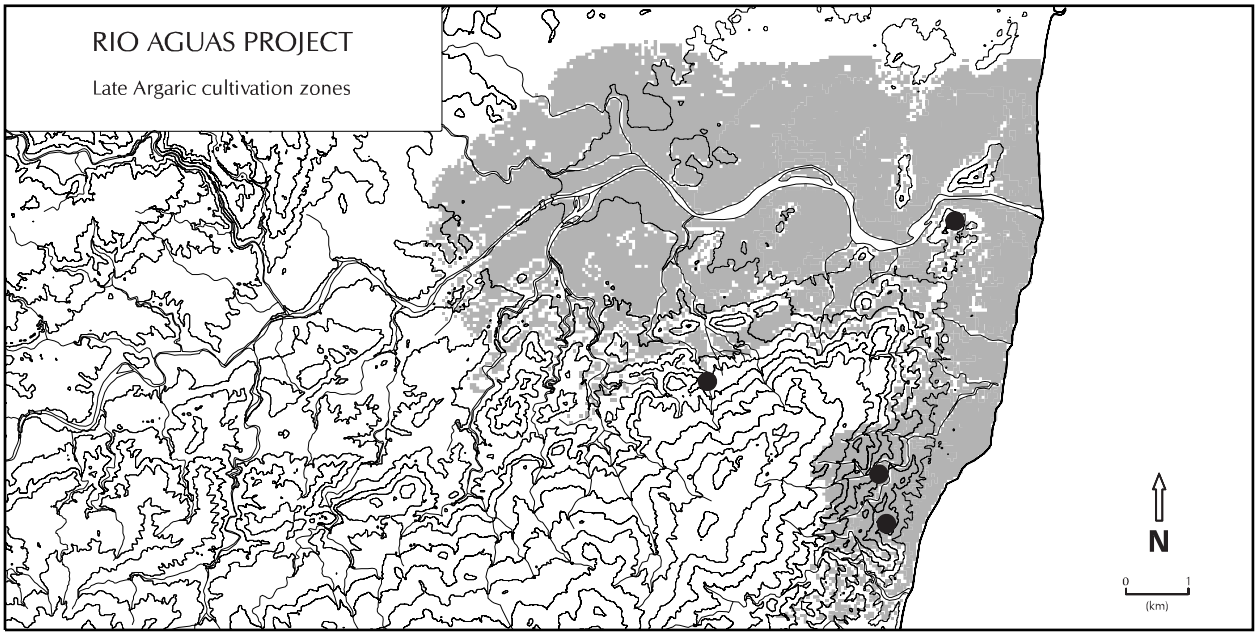


Figure 8.9. Hypothetical cultivation zones for the Late Argaric period.

LATE AND FINAL BRONZE AGE (1,550 - 600 BC; figure 8.10)

A drastic population decline coupled to a change in subsistence strategy toward combined *regadío* and *secano* leads to smaller cultivation zones, which are limited to areas relatively close to the Río Aguas and to the coastal zone. It is again questionable if the coastal zone was actually cultivated during this period given its generally low land use potential. From the Argaric period, only the sites of Gatas and Barranco de la Ciudad survive, while some new and smaller settlements are observed closer to the Río Aguas. From the archaeological evidence it is clear that the population continued to decline during this period; until the Roman Imperial period, other areas of southern Spain seem to have been more attractive for settlement, as is shown by the archaeological record (Chapman, 1991).

PHOENICIAN AND ROMAN REPUBLICAN PERIOD (600 BC - AD 0; figures 8.11 and 8.12)

Population reaches its lowest point during this period. During the Phoenician period (600-100 BC), only one settlement is found. This site is placed in an area very well suited for dry farming. In the Roman Republican period, two settlements are found, one near the Río Aguas where *regadío* could be applied, and one at the coast where *secano* will have been the best option.

ROMAN IMPERIAL PERIOD (AD 0 - 400; figure 8.13)

In the Roman Imperial period, a drastic increase in population is observed. The introduction of the Villa system leads to a very intensive agricultural exploitation of the Río Aguas valley. Most settlements seem to have been relying on dry farming rather than *regadío*, which confirms the evidence from the archaeological

record. Out of the eleven settlements, only two are found at somewhat more remote locations: Marina de la Torre in the mouth of the Río Aguas and Cerro del Picacho, which seems to have been chosen for its strategic position. No settlements are found that have extremely low attractivity indices, which is no surprise given the Roman policy of generating surplus food production. It is however clear from the modelling that this surplus production can not have been achieved with dry farming, as the complete Río Aguas valley should be taken into cultivation to feed the population of the settlements alone. This explains the suggested shift in this period towards *regadío*, made possible by the introduction of the first irrigation systems along the Río Aguas.

LATE ROMAN / VISIGOTHIC / BYZANTINE PERIOD (AD 400 - 700; figure 8.14)

During this period the population shows no drastic change, and many settlements continue their existence from the Roman Imperial period. The more decentralised occupation pattern is coupled to a further development of hydraulic infrastructure and irrigation in the valleys, as is also suggested by the dominance of wheat in the subsistence production. Surplus production was no longer maintained because of a lack of external demand, and the population may have reduced its agricultural activities to the most productive ones for its own subsistence needs. This is reflected in the cultivation zone mapping for this period: the area cultivated is limited and mainly found in areas with high land use potential.

OMEYA - CALIFAL PERIOD (AD 700 - 1200; figure 8.15)

After the Arab conquest of Spain, a new settlement pattern emerges in the area. While maintaining the application of *regadío*, dry farming is re-introduced in the area, probably as a consequence of external demand. The location of the settlements reflects this focus on dry farming. Only one exception is found: the site of Cerro de Inox is an Arab castle built high in the Sierra Cabrera. A beginning is made with the introduction of the complex hydraulic and terracing systems on the slopes of the Sierra Cabrera, which can still be observed today. This is however not reflected in the mapping of the cultivation zones, which are based on a strategy of predominantly dry farming. It is clear that most of the Río Aguas valley will have been cultivated during this period.

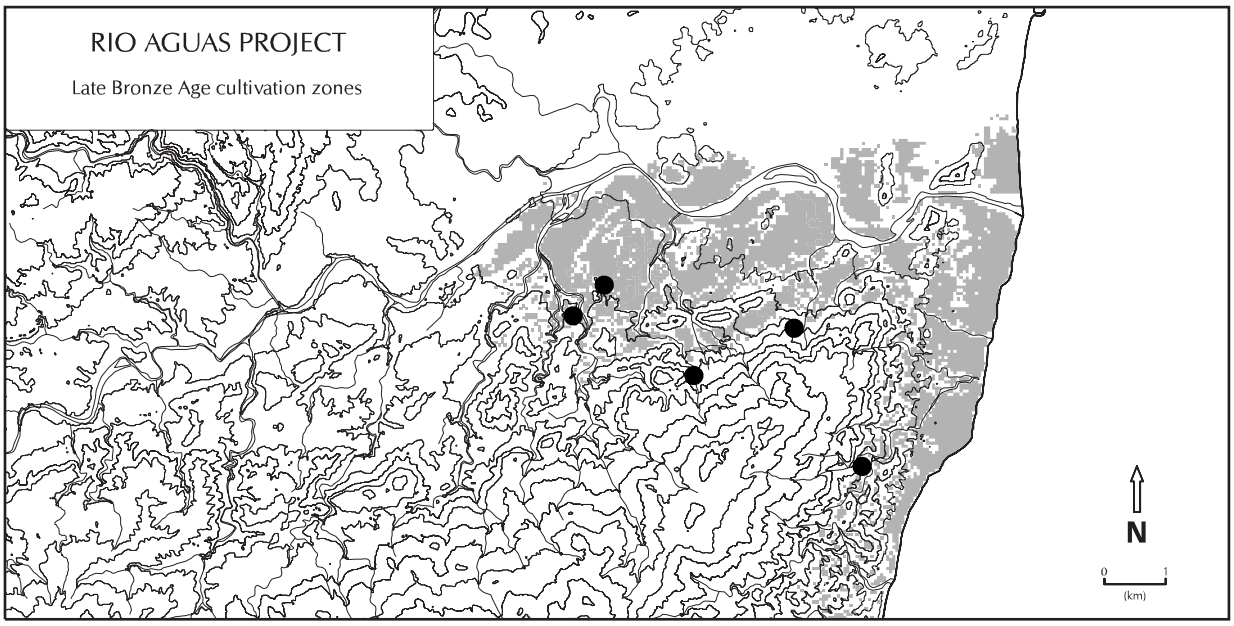


Figure 8.10. Hypothetical cultivation zones for the Late and Final Bronze Age.

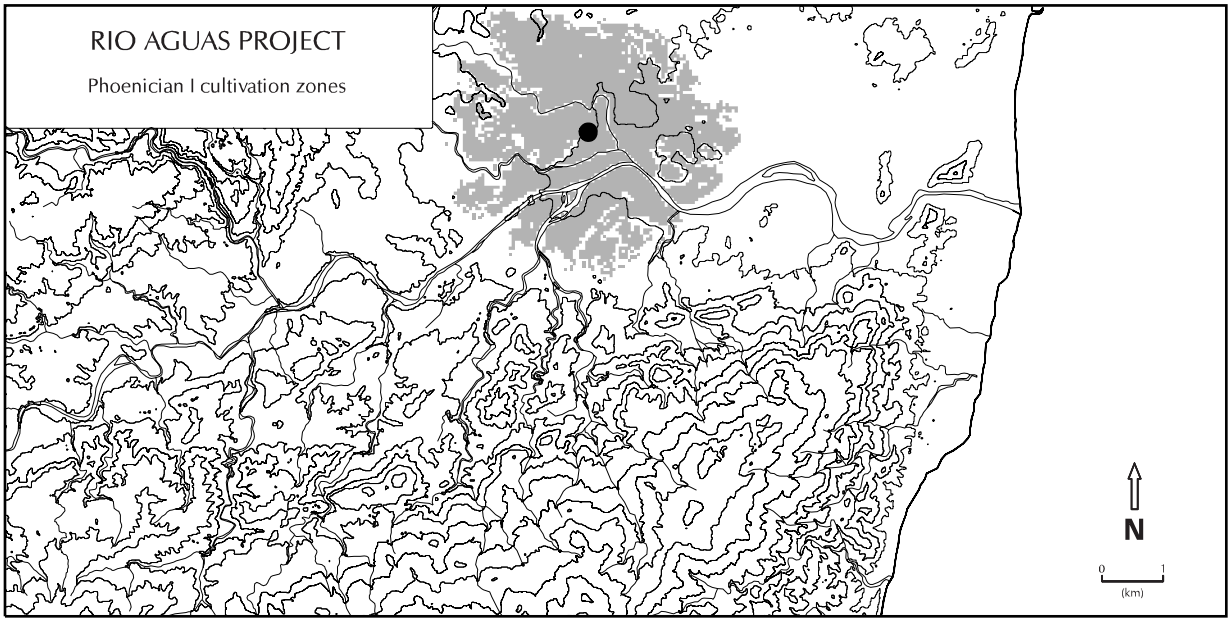


Figure 8.11. Hypothetical cultivation zones for the Phoenician period.

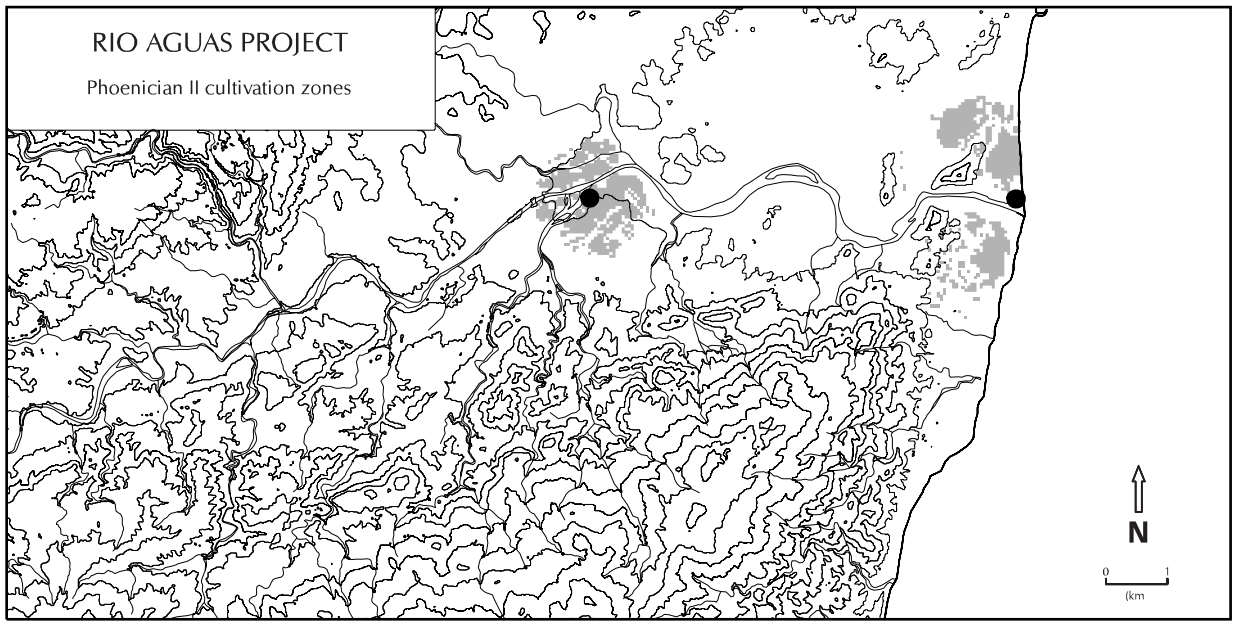


Figure 8.12. Hypothetical cultivation zones for the Roman Republican period.

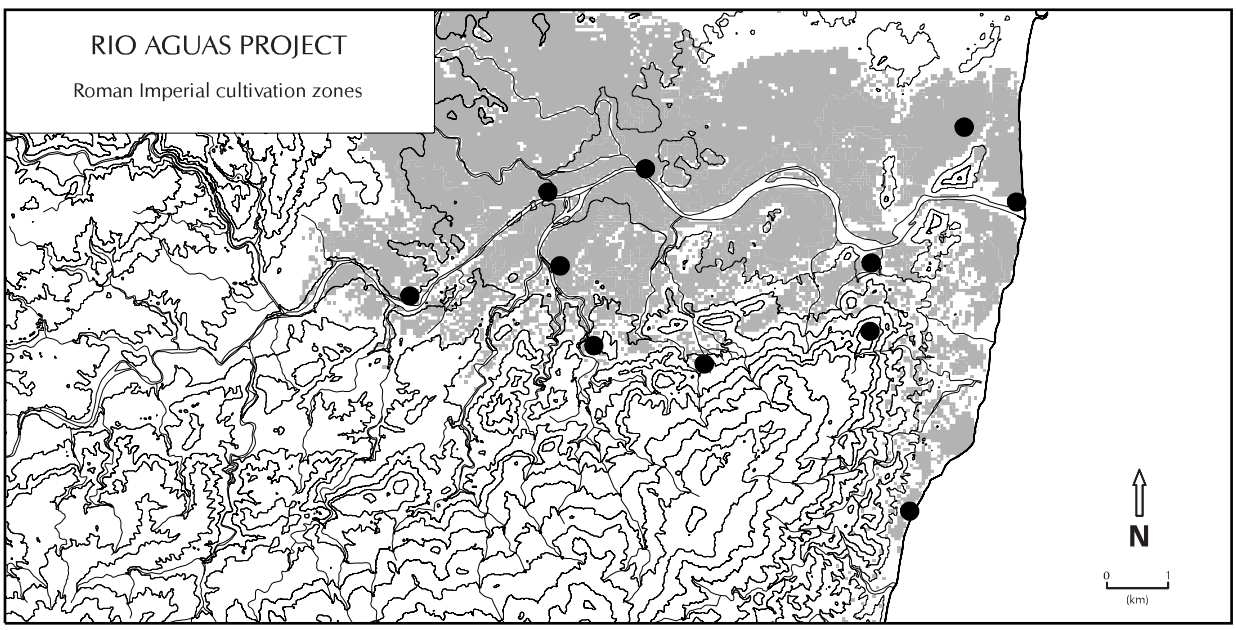


Figure 8.13. Hypothetical cultivation zones for the Roman Imperial period.

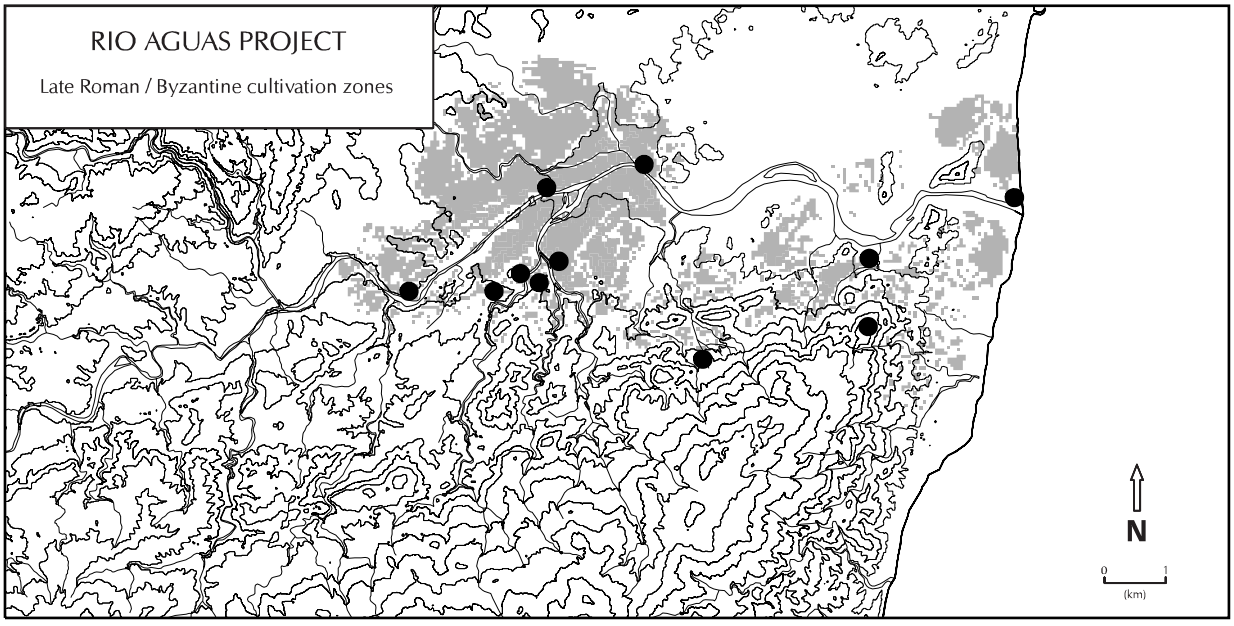


Figure 8.14. Hypothetical cultivation zones for the Late Roman / Visigothic / Byzantine period.

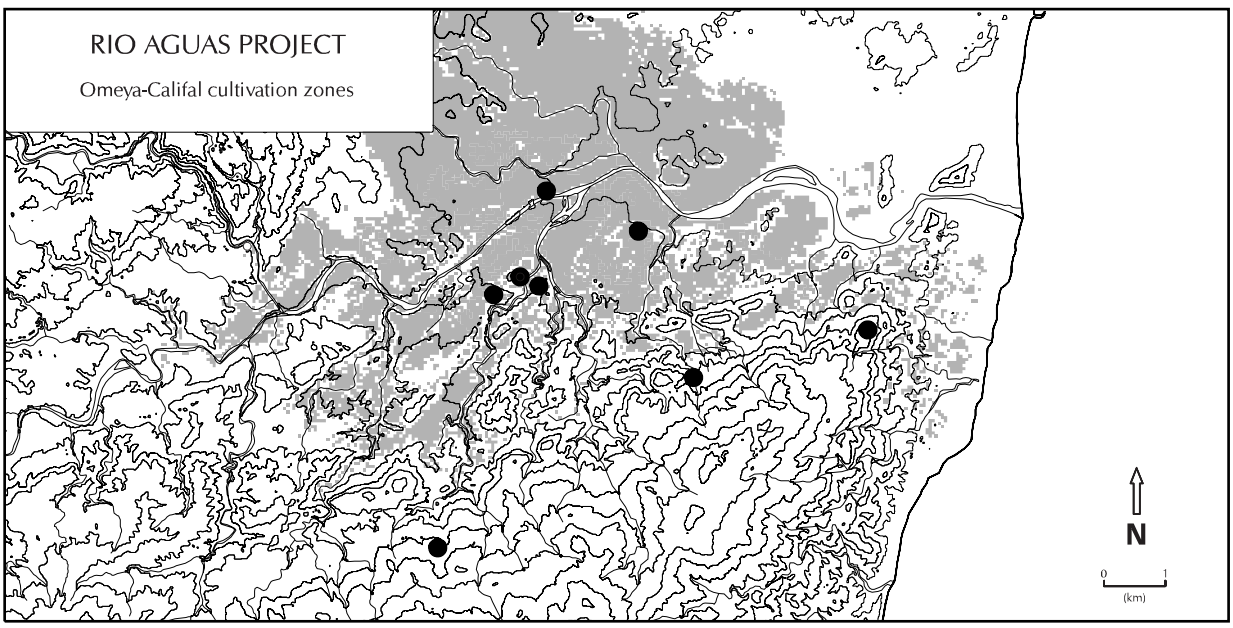


Figure 8.15. Hypothetical cultivation zones for the Omeya - Califal period.

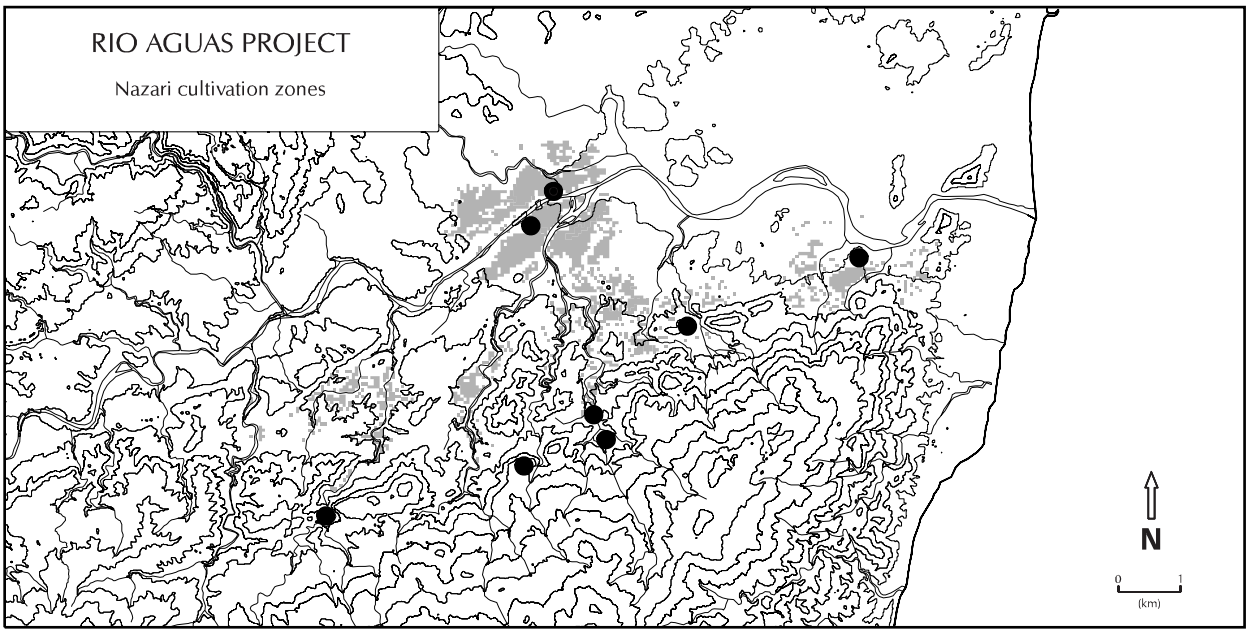


Figure 8.16. Hypothetical cultivation zones for the Nazarí period.

NAZARÍ PERIOD (AD 1200 - 1500; figure 8.16)

During this last phase of the Arab period, the settlement pattern has changed considerably. Two types of settlements can be observed: those located in the Río Aguas flood plain with good access to agricultural land, and those located on the slopes of the Sierra Cabrera, relying on the now fully developed terracing system for their agricultural production. Surplus production of cereals was reduced in favour of mulberry plantations (related to silk production) and olive trees. The mapping of the cultivation zones shows that in fact very small areas were necessary for subsistence production, and that the largest sites are found in the Río Aguas flood plain. Because of the independence of the settlements from high potential land, the mapping is not providing a very realistic picture of the possible cultivation zones.

8.9. CONCLUSIONS

The model discussed above provides a considerable amount of information on the possible development of the agricultural production pattern in the area in prehistory. At the regional level, it shows the possible impact of agriculture on the landscape for each time frame considered. The cultivation zone maps indicate which locations are the most probable production areas. The model assumptions do however not always conform to the actual situation. This is especially evident for the Neolithic period, when settlements did not exist simultaneously, and for the Nazarí period, when the technological innovation of terracing made it possible to grow crops almost regardless of the land use potential, so for these periods the cultivation zone mapping may not be reflecting the real situation. On the other hand, the model strongly supports the hypothesis of over-exploitation during the Late Argaric period, and points to the necessity of irrigation in order to obtain surplus production during the Roman period.

At the settlement level, the model indicates which sites may not have been relying on agriculture for their subsistence needs, and whether *regadío* or *secano* may have been the most probable option. The model also points to situations where land use conflicts between settlements may have arisen.

Finally, from the point of view of long term land degradation, a measure of the degree of exploitation of the landscape can be obtained by adding all the land use distribution maps into one single map (figure 8.17). From this exploitation intensity map, it can be concluded that the area around the mouth of the Rambla de Mofar will have been the most frequently used for agriculture. This is a 'high potential' area that is not very susceptible to land degradation. The area north of the Río Aguas around Las Alparatas however is not very frequently included in the cultivation zones, in spite of its relatively high potential. This may point to a lack of archaeological knowledge of this area located close to the town of Turre. Most other areas that show infrequent exploitation are classified as low potential land, with associated high degradation risks. The modelling therefore suggests that there is an upper threshold of available land that can be exploited without running into problems of land degradation. It is within these limits of possible exploitation that different solutions have been adopted through time to the question of agricultural subsistence production.

As a concluding remark, we may state that the GIS has proved to be an indispensable instrument for the modelling discussed here. It is however evident that the model will only serve its purpose if it provokes a dialogue with the archaeological and palaeo-environmental hypotheses and evidence involved. In this sense, the modelling presented here provides a working example of how GIS can be incorporated into a broader archaeological research framework.

ACKNOWLEDGEMENTS

This study was done in the framework of the Río Aguas Project, funded by the Directorate General XII of the Commission of the European Union under contract EV5V-CT94-0487, the Dirección General de Investigación Científica y Técnica del Ministerio de Educación y Ciencia, and the Comissionat per Universitats i Recerca de la Generalitat de Catalunya.. The authors would also like to acknowledge the co-operation of the colleagues involved in the Río Aguas Project, especially Pedro V. Castro and Ma. Encarna Sanahuja (Universitat Autònoma de Barcelona).

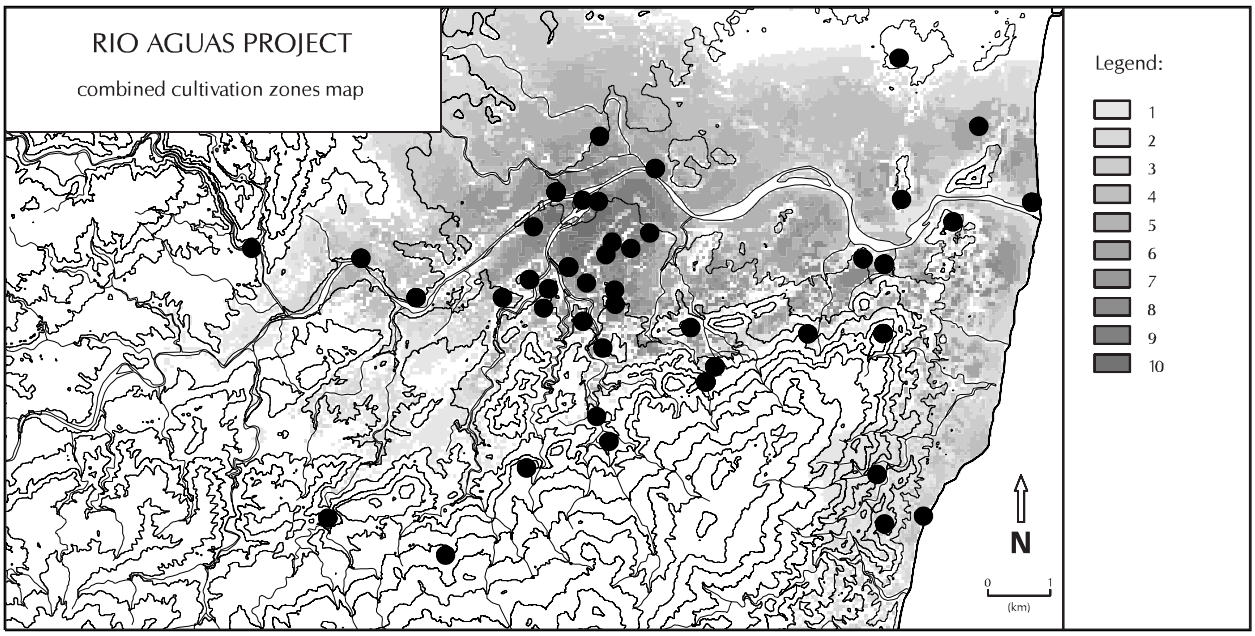


Figure 8.17. Combined cultivation zone map for all periods.

BIBLIOGRAPHY

Castro, P., Colomer, E., Courty, M. A., Fedoroff, N., Gili, S., González Marc, N. P., Jones, M. K., Lull, V., McGlade, J., Micó, R., Montón, S., Rihuete, C., Risch, R., Ruiz Parra, M., Sanahuja, M. E., and Tenas, M. (eds.), 1994. *Understanding the natural and anthropogenic causes of soil degradation and desertification in the Mediterranean Basin, Volume 2: Temporalities and desertification in the Vera Basin*. Commission of the European Union, Brussels.

Castro, P., Chapman, R., Gili, S., Lull, V., Micó, R., Rihuete, C., Risch, R. and Sanahuja, M. E. (eds.), 1996. *Río Aguas Project. Palaeoclimatic reconstruction and the dynamics of human settlement and land-use in the area of the middle Aguas (Almer¡a) of the south-east of the Iberian Peninsula.* Commission of the European Union, Brussels

Castro, P., Chapman, R., Gili, S., Lull, V., Micó, R., Rihuete, C., Risch, R., and Sanahuja, M. E. (eds.), 1997. *Proyecto Gatas 2. La dinámica arqueoecológica de la ocupación prehistórica* Consejería de Cultura de la Junta de Andalucía, Sevilla

Chapman, R. W., 1991. *La formación de las sociedades complejas. El sureste de la península ibérica en el marco del Mediterráneo occidental.* Barcelona.

Chisholm, M., 1962. *Rural settlement and land use: an essay in location*. Hutchinson University Library, London.

Clapham, A. J., Jones, M. K., Reed, J., and Tenas, M., 1997. 'Análisis carpológico del proyecto Gatas', in: Castro, P., Chapman, R., Gili, S., Lull, V., Micó, R., Rihuete, C., Risch, R., and Sanahuja, M. E. (eds.), 1997. *Proyecto Gatas 2. La dinámica arqueoecológica de la ocupación prehistórica* Consejería de Cultura de la Junta de Andalucía, Sevilla

Gilman, A. and J. Thornes, 1985. *Land use and prehistory in south-east Spain.* The London Research Series in Geography 8. George Allen & Unwin, London.

Gorenflo, L. J., and Gale, N., 1990. 'Mapping regional settlement in information space'. *Journal of Anthropological Archaeology*, 9:240-274

Leeuw, S. E. van der (ed.), 1994. *Understanding the natural and anthropogenic causes of soil degradation and desertification in the Mediterranean Basin.* Commission of the European Union, Brussels.

Lull, V., 1983. *La cultura de El Argar. Un modelo para el estudio de las formaciones económico-sociales prehistóricas.* Akal, Madrid

Ministerio de Agricultura, Pesca y Alimentación, 1982. *Evaluación de Recursos Agrarios: Mapa de Cultivos y Aprovechamientos 1:50.000. Garrucha (1015), Sorbas/Mojácar (1031/1032), Vera (1014).* Madrid.

Renfrew, C., 1972. *The Emergence of Civilisation. The Cyclades and the Aegean in the third millenium BC.* London.

Schowengerdt, R. A., 1983. *Techniques for Image Processing and Classification in Remote Sensing.* Orlando.

Schulte, L., 1996 ‘Morfogénesis Cuaternaria en el curso inferior del Río de Aguas (Cuenca de Vera, Provincia de Almería)’, in Grandal d’Anglade, A. and Pagés Valcarlos, J. (eds.): *IV Reunión de Geomorfología*. O Castro.

POSTSCRIPT TO CHAPTER 8

This chapter builds on an earlier paper (Verhagen *et al.*, 1995), which tried to demonstrate that GIS could do more than just represent geographical data in cartographic form, and only obtain statistical, descriptive information from those data. In fact, it was argued, GIS is the ideal tool to create and visualize hypotheses of human spatial behaviour. In this view, the formalization of these hypotheses will create a ‘dialogue’ between the ‘interpretative model’ and the archaeologists, confronting them with possible flaws in their way of reasoning. This is a well accepted approach to modelling outside archaeology: realism is not what is aimed for, but rather the isolation of the parameters of interest, and the extrapolation of (assumed) causal relationships, that can then be tested by using the available evidence. At the time, it was almost impossible to achieve more than very crude temporal modelling in GIS, and the model presented here is therefore rather simple, using time slices. Furthermore, the assumptions concerning demographic development are crude, to say the least. However, this is not the point of the modelling: realism is not intended. Rather, the model delimits a possibility space. If our assumptions on demography, diet and land suitability are correct, does this mean that Argaric society could still be self-sufficient? Can it be demonstrated that Roman agricultural production had to be done by means of irrigation? How do we explain the location of sites that have difficulty providing for themselves? By playing with various scenarios, in some cases a clear answer can be given, in other cases the verdict is still out.

The relevance of this type of land use reconstructions to predictive modelling is found in the freedom of scenario building. By creating and comparing various scenarios with the available data, it is possible to arrive at the most probable scenario – and this will be the scenario that provides the best predictive model. Given the development of software and hardware, the effort of recalculating these scenarios (which in the late 1990s would still take days), has considerably decreased. Such an approach is therefore now a realistic option, whereas it was only a dream ten years ago.

ADDITIONAL REFERENCE

Verhagen, P., J. McGlade, S. Gili and R. Risch, 1995. 'Some Criteria for Modelling Socio-Economic Activities in the Bronze Age of south-east Spain', in: Lock, G. & Stančič, Z. (eds.): *GIS and Archaeology: A European Perspective*. Taylor and Francis, London, pp. 187-209.

CHAPTER 9 Some considerations on the use of archaeological land evaluation[1]

9.1. INTRODUCTION

Land evaluation can be useful as a technique to gain insight into the possibilities and limitations of agricultural production of the past. In general, it will be better suited for studies that cover large areas and aim for general conclusions on agricultural production at the landscape scale, as this is also the scope of the original *Framework for Land Evaluation* that was published by the FAO (1976). The number of applications found in archaeological literature however is limited (Boerma, 1989; Kamermans, 1993; Finke *et al.*, 1994)[2]. These applications can all be characterized as examples of qualitative land evaluation, which is an essentially deductive technique that leans heavily on multi-criteria analysis methods. As such, it has recently also received some attention as a tool to explain site location at a regional level (Kamermans, 2000), and a few more applications can be found which use deductive modelling for the explanation of site location without direct reference to land evaluation (e.g. Chadwick, 1978; Doorn, 1993; Dalla Bona, 1994). In general however, the impact of land evaluation and related deductive methods on archaeological practice has been limited.

A important reason for this may be the lack of a theoretical framework for the application of land evaluation in archaeology. Deductive modelling in archaeology starts with the construction of hypotheses on the behaviour of prehistoric people and the dynamics of the environment in the past, and therefore needs a strong theoretical basis. Modern land evaluation is based on the assumption of economic optimization, which is a questionable hypothesis for most prehistoric societies, and does not incorporate the temporal effects that are so essential to archaeology. In order to carry out an archaeological land evaluation that makes sense, the role of environmental change, technological development and the human perception of land suitability are of primary concern, and need to be incorporated into the modelling. If these issues are downplayed or ignored, land evaluation will produce crude models of prehistoric agricultural potential, which may even be based on questionable archaeological and environmental evidence.

The current paper will try to highlight the issues of environmental change, technological development en human perception in the context of prehistoric land evaluation. This will be done using some of the results of the EU-funded Archaeomedes I (van der Leeuw, 1998) and Rio Aguas (Castro *et al.*, 1998) projects, which have been carried out in the years 1992-1996, and in which the author has been involved in the reconstruction and analysis of prehistoric land use, and its consequences for long term land degradation in the Mediterranean.

---

[1] This paper also appeared in Attema, P., G.-J. Burgers, E. van Joolen, M. van Leusen and B. Mater (eds.), 2002: *New Developments in Italian Landscape Archaeology*. British Archaeological Reports, International Series 1091. Archaeopress, Oxford, pp. 200-204.

[2] obviously, this only refers to papers concerning formal land evaluation; the literature on site catchment analysis and the reconstruction of agricultural production zones in archaeology is quite substantial

9.2. ENVIRONMENTAL CHANGE AND ITS CONSEQUENCES FOR LAND SUITABILITY

Land qualities, and therefore land suitability will change in time *and* space as a consequence of environmental change, either from natural causes or human impact. When trying to assess the possible effects of environmental change on land suitability, the most important factors to be reconstructed are vegetation, climate and hydraulic regime, as these determine the geomorphic and hydrologic dynamics of a region. Without an understanding of these three basic elements, it will be very hard to assess possible changes in land suitability over time. The information necessary is usually only available through palaeo-botanical and palaeo-pedological research, and as such the level at which environmental change can be incorporated into the modelling is highly dependent on the amount of accessible palaeo-environmental data.

THE TEMPORAL DIMENSION

It will be evident that environmental change may affect land qualities like soil moisture content, fertility and soil depth through time. However, environmental change does not necessarily have an immediate effect on the agricultural potential of a region. For example, in the French Rhône Valley (van der Leeuw, 1998), large scale deforestation is evident as early as the 3rd or 4th century BC, but widespread erosion did not occur until *after* the Roman period. The delay is presumably coupled to changing climatic conditions in the Early Middle Ages, when a more humid phase triggered an erosion process that was waiting to happen. A time delay can be observed between vegetation degradation and geomorphic consequences; vegetation is more susceptible to human impact, whereas geomorphology reacts primarily to climatic changes, notably in humidity and seasonal contrast. When human pressure, vegetation degradation and climate change occur simultaneously, the response is almost immediate and dramatic, but when they are out of phase, delays will occur.

It is also important to note that environmental crises are usually short-lived; periods of stability are longer lasting, but poorly documented in the pedological record. However, the crises may be the triggers that cause the whole geomorphic system to search for a new state of equilibrium[3], and as such can have serious consequences for land suitability.

SPATIAL EFFECTS

Apart from the temporal delays mentioned, spatial effects should be considered as well. In the case of a phase of increased erosion, its consequences for land suitability highly depend on the position of an area in the landscape. On lower slopes, aggradation may compensate degradation, and the suitability of the land for agriculture may not change very much under more dynamic geomorphic conditions. Low-lying areas however may be subjected to regular flooding when an increased sediment deposition blocks the natural (or artificial) drainage sytems. And to give an additional, somewhat counter-intuitive example from the very recent past: in northern Epirus (Greece), massive forest regeneration can be observed over the past two or three decades, because the hillsides are no longer used for sheep and goat grazing. This development has led to a decreasing availability of drinking water in the valleys, the water being used by the trees before it can reach the valley bottom.

[3] Bintliff (2002) refers to this phenomenon as ‘punctuated equilibrium’

APPLICATION TO LAND EVALUATION

Unfortunately, the effects of environmental change are easily overlooked in prehistoric land evaluation, as the FAO framework itself was never meant to be used in a prehistoric context using palaeo-environmental reconstructions. For example, Finke *et al.* (1994) use modern day meteorological data to assess soil moisture availability and evapotranspiration deficits for a land evaluation of the Gubbio Basin (Italy) in the Bronze Age. Similarly, they use modern soil maps to assess a number of soil characteristics, like soil drainage and rooting depth, and assess flood hazard from the modern day distribution of alluvial deposits. This is not to say they have been wrong in doing so; however, the authors do not make an argument to defend the use of such data for a Bronze Age landscape.

An argument can be made that as long as the effects of environmental change are occurring uniformly over the whole area, the *relative* suitability of the land will not change, although in absolute terms conditions may be improving or deteriorating for a certain type of land use. However, as soon as asymmetries can be observed in the response of certain areas to environmental change, it is evident that there will be changes in relative suitability as well. The examples given above show that these asymmetries will occur, both in time and space, and should be analyzed in order to produce a prehistoric land evaluation that makes sense.

This does not necessarily imply that the environmental reconstructions need to be very detailed. It does however imply that the FAO framework should not be used as a rigid standard that should be applied under all circumstances. The FAO land evaluation system poses strict thresholds on land qualities in order to qualify for inclusion in a suitability class. Some of these thresholds, like those for soil fertility, can only be determined by means of laboratory analysis. Even when no chemical analyses are needed, the definition of the thresholds may be problematic for an archaeological application. Flood hazard for example is classified into four classes:

F0	no floods
F1	infrequent floods (< 5 times per 10 years)
F2	frequent floods (> 5 times per 10 years)
F3	regular floods (every year)

It is obvious that even with a well-documented geological record, classification of areas in especially classes F1 or F2 will be impossible to do in a prehistoric context; apart from that, the importance of the periodicity of floods may have been different in the past. Much more interesting than knowing exactly how often an area was flooded may be the question if the area was more or less susceptible to floods during certain periods, or whether floods may have been occurring in other seasons.

9.3. TECHNOLOGICAL DEVELOPMENT: HYDRAULIC INFRASTRUCTURE

The changing level of technological development is a factor that influenced the possibilities of prehistoric societies to take lands into cultivation. Although technological development in agriculture has occurred in many areas (e.g. the development of new plough types, the introduction of animal traction and in modern times the use of fertilizer), probably one of the most significant developments has been the introduction and sophistication of hydraulic infrastructure in areas where an excess or shortage of water was experienced.

WATER EXCESS

In areas where excess of water was a problem, like the Tricastin area in the Middle Rhône Valley, the Romans developed an extensive drainage infrastructure in order to exploit the humid floodplain of the Rhône (van der Leeuw, 1998). The system was designed to manage the hydrology of a very large area, and structured the pattern of Roman colonisation of the area. The available evidence indicates that the system was highly vulnerable to changing environmental circumstances. A modest increase in precipitation resulted in notable erosion and sedimentation problems, causing blocking of the drainage system. The fact that the area was not colonised very rapidly implied that when climatic conditions changed the area had not been settled fully, and manpower was lacking to maintain the drainage system at the level necessary to prevent flooding.

Ultimately, this means that even though the Romans had the technological expertise to take the land into cultivation, the implementation of the drainage system was highly dependent on the availability of manpower and willingness to invest in the colonisation of the area. There is strong parallel with modern land evaluation here: it is common practice to determine land suitabilities conditional on the availability of certain technological improvements, like a drainage system. In the end however economic and social circumstances determine whether these improvements will be made, and whether they can be maintained.

WATER SHORTAGE: THE RIO AGUAS EXAMPLE

In the south-east of Spain, it is generally thought that the combination of human impact and climatic factors has led to a general aridification of the area since the Neolithic. As a consequence, the development hydraulic infrastructure for irrigation has been very important for agricultural production through time. In order to cope with drier conditions, (pre-)historic societies have developed and adapted the water distribution system in the area again and again. For the Rio Aguas Project, an attempt was made to analyse this development of food production with changing technological innovation (Castro *et al.*, 1998; Verhagen *et al.*, 1999).

In essence, three different types of irrigation can be distinguished in the area:

- natural irrigation by inundation (*regadío*) of areas close to the river bed of the Rio Aguas, available to all (pre)historic societies;
- irrigation of the rest of the floodplain by means of canals, probably introduced in the Roman Imperial period; and
- irrigation of the lower hill slopes by means of a combination of water conduits (known as *acequias*) and terracing, introduced in the Arab period.

As a first step, a basic land evaluation was carried of the area for each of the three irrigation types. The distribution of land use in 1978 (when agriculture in the area was not yet modernized to a large degree) was used to determine where each of the three irrigation types was found in the recent past, and the results of this analysis were extrapolated to the whole area, as not all of it was used for agriculture in recent times. Not surprisingly, the most probable areas for natural irrigation were found close to the river bed of the Rio Aguas, whereas areas suitable for irrigation of type b) and c) were found at increasingly large distances from the river. The suitability maps were then combined with data on settlement distribution and estimates of demographic development through (pre)history in order to model hypothetic agricultural production zones for each period under a hypothesis of self-sufficiency. The models produced were used to judge if potential food production problems might have occurred, and if irrigation could have been applied to counteract these problems.

From archaeological evidence it is clear that during the Neolithic and Chalcolithic no artificial irrigation was applied; the modelled pattern of cultivation zones conformed to this evidence, as settlements are found close to water resources and needed relatively modest amounts of land for crop cultivation.

However, a sharp population increase during the Early and Middle Bronze Age (the Argaric period) meant that much larger areas were needed for cultivation. The demographic rise was coupled to a period of drier and hotter climatic conditions, and these combined factors evidently led to agricultural production problems. Argaric society seems to have tried to counteract the problem by introducing a monoculture of barley, which is a highly drought resistant crop, and could therefore be grown in areas not well suited for other crops. However, this strategy seems to have led to nutritional problems, large scale deforestation in the river valley and possibly the eventual collapse of Argaric society when it could no longer cope with the environmental problems. It is not clear whether artificial irrigation was technologically speaking an option to the Argaric people, but it is clearly a strategy that for some reason was not pursued.

Following a gradual but steady population decline, it was not until the the Roman Imperial period that the area was again used for extensive agriculture. The modelling showed that the most suitable soils should almost all be taken into cultivation in order to feed the population. However, it is well known that surplus food

production was furthered by the Roman Empire. This implies that artifical irrigation must have been applied in order to obtain the higher production levels needed.

During the Arab period, a complex terrace and hydraulic system was introduced, in order to be able to settle the mountain slopes that were previously unavailable for settlement. The system seems to have worked satisfactorily during the later Arab period as well, allowing for sufficient food production for the settlements as well as surplus production of olives and (silk production related) mulberry trees, and at the same time reducing environmental degradation. It is interesting to observe that after the Spanish *reconquista* the hydraulic infrastructure is abandoned because of 'insufficient productivity' according to contemporary sources. Again this points to the fact that the use of technological developments is highly dependent on social and economic circumstances.

9.4. THE HUMAN PERCEPTION OF SUITABILITY

Farmers in the Atlantic zone of Costa Rica, when asked if they could distinguish between good and bad soils, answered that one could easily recognize the good, black soils, suitable for almost any crop and the bad, red soils, that were only suitable for grazing. In essence, they applied a very simple form of land evaluation; formalized land evaluation methods are basically just tools to express the modern perception of land suitability in an unambiguous way. In the past, farmers will have judged the quality of land by using their own experience on what was good or bad for a certain type of land use. These judgments may have been different from our modern way of thinking about land quality, as different factors may have been involved, such as the fact if surplus production was aimed at or not, the level of labour input and technological investment needed, and the necessity of population centres to be close to the zones of food production. Apart from that, their understanding of crop requirements may have been different from ours.

THE ROMAN PERCEPTION OF LAND SUITABILITY

A glimpse of the way people thought about land suitability in the past can be obtained from Roman agronomical texts. Favory *et al.* (1995) have analysed the texts of Columella and Pliny the Elder in order to see if the Romans' view of agricultural suitability was very different from modern ones. One thing that is immediately evident is the emphasis placed on light, easily workable soils. Columella distinguishes four types of soil:

- rich and light soils, easy to work, giving the highest return with the least effort
- rich and compact soils, that require hard labour but compensate this with good returns
- humid soils, that do not need much labour, and can still produce a crop
- dry, compact and poor soils, that are difficult to work and do not yield a good return

This very basic description formed the basis of the Roman land taxation system, as has been demonstrated for the Tricastin area in the Middle Rhône Valley (Berger *et al.*, 1997; van der Leeuw, 1998).

Pliny the Elder, in his Natural History, goes far beyond Columella's soil classification. In his encyclopedic text, he describes for many crops what their requirements are, and the characteristics of many soil types, including the best crops to plant. The text does have errors and omissions in it, but it proved possible to compare the soil descriptions provided by Pliny with the categories distinguished on modern soil

maps. In this way, it is possible to perform a land evaluation more or less like Pliny would have done it, had he had soil maps available. One of the examples of the way Pliny described soils is given here, translated from Favory *et al.* (1995):

Humidus ager; Locum humidiore; Humidum solum

translated as	humid soil
most probable soil type	recent alluvial soils, sandy and humid but not hydromorphic
effect on crops	almond trees become sterile or even die
crops	emmer, asparagus, olives, elm trees; almond trees and alfalfa to be avoided
advice	dig up the soil for cultivation of asparagus; market garden culture
hydrology	close to water
to be recognised by	vicinity of water; vegetation type
topography	plains
vegetation	grassland
comparable soils	rich soils (not suited for almonds), deep soils (asparagus), dry soils (elm trees), grassland soils (not suited for alfalfa), very or slightly humid soils, clayey soils (olives)
opposed soils	dry soils on hillsides (different cultivation for elm trees), irrigated dry soils (well suited for alfalfa), warm and hard clay soils (well suited for almond trees), dry clayey soils (different planting season for olives)

From the comparison of Pliny's soil descriptions and modern agronomic texts it can be concluded that the Romans were not overly concerned with the concept of soil fertility. Nowadays, fertile soils are associated with the profitable cultivation of fruit trees, but the Romans thought fertile soils were those suited for the cultivation of the traditional crops of wheat, olives and vines, which formed the basis of the Roman subsistence economy. On the other hand, the Romans placed a much stronger emphasis than modern agronomers on the workability of soils, not surprising in an era when manual labour was still very important in order to successfully grow any kind of crop.

Crop requirements as recognized by Pliny may also be a little bit different from modern opinions, as is illustrated by this example:

Asparagus
Requirements according to modern agronomers: Needs a sandy soil, cool but well drained, neither humid nor calcareous. Resistant to winter frost, but sensitive to spring frost. Can be cultivated on almost any location, except on high ground. Needs deep ploughing and manuring .
Requirements according to Pliny: Sow it in a humid or deep soil that has been dug up. Saturate it with manure. Weed often. The best suited soils for the cultivation of asparagus are the soils of the gardens of Ravenna.

The Roman and modern view agree on the need for a deep soil that is manured, but Pliny insists on a humid soil, whereas modern agronomers don't. Apart from that, the number of criteria used by modern agronomers is much larger and includes requirements on climatic conditions, chemical composition and texture of the soil (note however that Pliny's definition of humid soils probably already implies a sandy texture). A land evaluation for asparagus based on modern criteria will probably result in a more restricted area

suited for asparagus cultivation, and will possibly classify different soil types as suitable than an evaluation based on Pliny's criteria.

9.5. CONCLUSIONS

Land evaluation can play a role in the study of (pre)historic societies as a method to establish the possibility space available for subsistence production. However, from the examples given in this paper it will be clear that a modern view of land evaluation will not be suited to answer archaeological questions. Even though the basics of crop requirements and soil characteristics are the same now and in the past, both environmental and human factors may radically change the outcome of an archaeological land evaluation, as the result of such a land evaluation should be to reconstruct the perception of (pre)historic societies with regard to land suitability, rather than the modern perception of it. It is the interplay of environmental factors, technological development and social and economic structures which determine whether prehistoric societies will have regarded certain areas suitable for crop cultivation (or other uses). It may be impossible to exactly reproduce the perception of (pre)historic people with regard to the suitability of soils, but it is hoped that the examples given here at least give an indication of how to pursue land evaluation in an archaeological context.

BIBLIOGRAPHY

Berger, J.-F., F. Favory, T. Odiot and M.-P. Zannier, 1997. 'Pédologie et agrologie antique dans le Tricastin central (Drôme-Vaucluse), d'après les textes agronomiques et épigraphiques latins et les données géoarchéologiques', in: J.P. Bravard, G. Chouquer and J. Burnouf (eds.), *La dynamique des paysages protohistoriques, antiques, médiévaux et modernes. XVIIe Rencontres Internationales d'Archéologie et d'Histoire d'Antibes.* Editions APCDA, Sophia-Antipolis, pp. 127-154.

Bintliff, J., 2002. 'Time, process and catastrophism in the study of Mediterranean alluvial history: a review'. *World Archaeology* 33:417-435.

Boerma, J.A.K., 1989. 'Land evaluation in prehistoric perspective: some observations', in: Haex, O.C.M., H.H. Curvers and P.M.M.G. Akkermans (eds.), *To the Euphrates and beyond.* Rotterdam, pp. 17-28.

Castro, P.V., R.W. Chapman, S. Gili, V. Lull, R. Micò, S. Montón, C. Rihuete, R. Risch and M.E. Sanahuja-Yll, 1998. *Aguas Project: Palaeoclimatic reconstruction and the dynamics of human settlement and land-use in the area of the middle Aguas (Almería), in the south-east of the Iberian Peninsula. Research results.* Office for Official Publications of the European Communities, Luxemburg.

Chadwick, A.J., 1978. 'A computer simulation of Mycenaean settlement', in: Hodder, I. (ed.), *Simulation studies in archaeology*. Cambridge University Press, Cambridge, pp. 47-57.

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

Doorn, P.K., 1993. 'Geographical Location and Interaction Models and the Reconstruction of Historical Settlement and Communication: The Example of Aetolia, Central Greece.' *Historical Social Research*, 18:22-35.

FAO, 1976. *A Framework for Land Evaluation.* ILRI Publication 22/FAO Soils Bulletin 32, Rome.

Favory, F., J.-J. Girardot and M.-P. Zannier, 1995. 'La perception des sols et des plantes chez les agronomes romains', in: van der Leeuw, S.E. (ed.), *The Archaeomedes Project: Understanding the natural and anthropogenic causes of soil degradation and desertification in the Mediterranean Basin. Volume 3: Dégradation et impact humain dans la moyenne et basse vallée du Rhône dans l'Antiquité (part II).* University of Cambridge, Cambridge, pp. 73-114.

Finke, P., J. Hardink, J. Sevink, R. Sewuster and S. Stoddart, 1994. 'The dissection of a Bronze and Early Iron Age landscape', in: C. Malone and S. Stoddart (eds.), *Territory, Time and State. The archaeological development of the Gubbio Basin.* Cambridge University Press, Cambridge.

Kamermans, H., 1993. *Archeologie en landevaluatie in de Agro Pontino (Lazio, Italië).* Universiteit van Amsterdam. Amsterdam. PhD Thesis.

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

Leeuw, S.E. van der (ed.), 1998. *The Archaeomedes Project: Understanding the natural and anthropogenic causes of soil degradation and desertification in the Mediterranean Basin. Research Results.* Office for Official Publications of the European Communities, Luxemburg.

Verhagen, P., S. Gili, R. Micó and R. Risch, 1999. 'Modelling prehistoric land use distribution in the Rio Aguas valley (province of Almería, S.E. Spain), in: L. Dingwall, S. Exon, V. Gaffney, S. Lafflin and M. van Leusen (eds.), *Archaeology in the Age of the Internet – CAA97. Computer Applications and Quantitative Methods in Archaeology 25th Anniversary Conference, University of Birmingham.* British Archaeological Reports, International Series 750. Archaeopress, Oxford. CD-ROM.

POSTSCRIPT TO CHAPTER 9

Land evaluation never really caught on as a method to create archaeological predictive models. Some of the reasons are given in this paper: in order to do it right, you need to have quite a lot of good archaeological and palaeo-environmental data. Van Joolen (2003) has tried to develop land evaluation further for archaeological purposes, and her thesis shows just how many parameters and considerations play a role in developing a working, multi-period archaeological land evaluation system. The approach of trying to determine the possibility space for agricultural production is of course not new (see also chapter 8), and also stands at the basis of the less formal expert judgment models that were discussed in chapter 4. These methods only make a very general assessment of productivity to select the areas best suited both for agriculture, and hence for human settlement. So the question is: when can land evaluation, being a more time consuming technique, contribute sufficiently to predictive modelling in order justify its application? Basically, we don't know. The amount of data collection needed is substantial (see also Kamermans and Rensink, 1999), and the advantages of using a formalised land evaluation approach (it is a generic technique, that applies an explanatory framework and is easily falsifiable), are also the advantages of any other formalised deductive modelling technique, like the models made by Whitley (2005) or Peeters (2005). It can however be argued that land evaluation has the additional advantage of using only measurable characteristics of the landscape. In the context of what has been discussed in this paper however, it is clear that this claim is built on thin ice. Many of the parameters needed can not be measured with certainty for the archaeological period(s) of interest. So, clearly, what is needed, is some form of cost-benefit analysis. Without comparative tests between methods and techniques, we will never know if it is worth the trouble investing time in very detailed palaeo-geographic reconstructions, or if it suffices to continue making models using only very general notions of human locational behaviour.

ADDITIONAL REFERENCES

Joolen, E. van, 2003. *Archaeological land evaluation: a reconstruction of the suitability of ancient landscapes for various land uses in Italy focused on the first millennium BC.* Rijksuniversiteit Groningen, Groningen. PhD thesis.

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. Archaeopress, Oxford. CD-ROM.

Peeters, H., 2005. 'The Forager's Pendulum: Mesolithic-Neolithic landscape dynamics, land-use variability and the spatio-temporal resolution of predictive models in archaeological heritage management', in: M. van Leusen and H.

Kamermans (eds), *Predictive Modelling for Archaeological Heritage Management: A Research Agenda.* Nederlandse Archeologische Rapporten 29. Rijksdienst voor het Oudheidkundig Bodemonderzoek, Amersfoort, pp. 149-168.

Whitley, T.G., 2005. '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. Rijksdienst voor het Oudheidkundig Bodemonderzoek, Amersfoort, pp. 123-138.

CHAPTER 10 First thoughts on the incorporation of cultural variables into predictive modelling[1]

*Philip Verhagen, Hans Kamermans[2], Martijn van Leusen[3], Jos Deeben[4], Daan Hallewas[4] and Paul Zoetbrood[4]*INTRODUCTION

Predictive modelling is a technique used to predict archaeological site locations on the basis of observed patterns and/or assumptions about human behaviour (Kohler and Parker, 1986; Kvamme 1988; 1990). It was initially developed in the USA in the late 1970s and early 1980s where it evolved from governmental land management projects and is still regularly applied in cultural resources management. In the Netherlands, predictive modelling plays an important role in the decision making process for planning schemes on a municipal, provincial and national level.

However, in many other countries predictive modelling is far from being an accepted tool for archaeological heritage management (AHM), and even where it is used regularly, criticism is not uncommon (see e.g. Ebert, 2000; Whitley, in press; van Leusen *et al.*, 2002). Much of this criticism is related to the uncritical application of so-called 'inductive' modelling techniques, in which the archaeological data set is used to obtain statistical correlations between the location of archaeological sites and environmental variables such as soil type, slope or distance to water. The performance of these models is in general not very good, partly because of the use of inappropriate statistical techniques, but mainly because of the biased nature of many archaeological data sets and the emphasis on environmental factors, which are easier to model than the more intangible social and cultural factors.

Wheatley (2003) even states that, as predictive modelling doesn't work very well, it shouldn't be used at all: "Archaeology should really face up to the possibility that useful, correlative predictive modelling will never work because archaeological landscapes are too complex or, to put it another way, too interesting". His argument is mainly directed against the use of biased archaeological data sets, that will lead to the development of biased models that will in turn inevitably produce a positive feedback loop of even more biased data sets, as it is common practice to spend funds for AHM on the areas of 'high archaeological value'. These areas will become better and better known, whereas the areas that are designated a 'low value' on the predictive map will largely be ignored in (commercial) archaeological research.

Verhagen (in press) however shows that the creation of biased data sets is not just a problem of predictive modelling, but a more general characteristic of the way in which archaeological data is collected. Most of the archaeological prospection done is not taking into account statistical sampling theory, and it can be suspected that many survey projects do not even have a strong archaeological hypothesis in mind. We believe that predictive modelling can serve as a means to make explicit the assumptions that are often made concerning the location preferences of prehistoric people. A predictive model should be based on a theory of site location preferences, that can be quantified and tested against (unbiased) archaeological data sets (see also

---

[1] This paper was presented by Hans Kamermans at the CAA 2004 conference, held from 13-17 April 2004 in Prato, Italy, and will be published in its proceedings. The text of this paper was largely prepared by me in co-operation with Hans Kamermans and Martijn van Leusen, but as it is part of the research done for the BBO-project 'Predictive Modelling', the other participants in the project are given credit as co-authors.

[2] Faculty of Archaeology, Leiden University

[3] Institute of Archaeology, Groningen University

[4] Rijksdienst voor het Oudheidkundig Bodemonderzoek, Amersfoort

Whitley, in press). It is clear that the cultural component of these theories is at the moment virtually absent in predictive modelling practice. This paper intends to show that it is not impossible to include these variables into predictive modelling, and this will hopefully lead to further research into this subject.

10.2. PREDICTIVE MODELLING AND ENVIRONMENTAL DETERMINISM

The practice of predictive modelling for AHM is, at the moment, environmental deterministic in outlook and design. The predominant use of environmental input variables as archaeological site predictors, such as soil type, groundwater table, distance to open water and the like, has however been criticized on a number of occasions in academic literature (e.g. Wheatley 1993; 1996a; 2003; Gaffney and van Leusen, 1995). The problems associated with environmentally based predictive modelling (van Leusen *et al.*, 2002) can be summarized as follows:

- archaeological theorists reject an understanding of past human behaviour in purely ecological/economical terms, and argue that social and cognitive factors determine this behaviour to a large extent, and should therefore be additional predictors for the presence and nature of archaeological remains;
- the maximum gain (a measurement of the degree of effectiveness of the predictive archaeological model over a 'by chance' model) of current predictive models seems to be about 70% (Ebert, 2000; Wheatley, 2003), which implies that a significant proportion of archaeological site locations cannot be predicted using purely environmental datasets; therefore, models based on environmental factors alone cannot be adequate tools for the prediction of archaeological site location.
- unfortunately, social and cognitive factors seem to be difficult to model, and have so far only be studied for a very limited range of questions, based on very specialised data sets (mostly relating to the ritual prehistoric landscapes of Wessex in England; e.g. Wheatley 1995; 1996b).

The American archaeologist Timothy Kohler observed this as early as 1988. "Why are the social, political, and even cognitive/religious factors that virtually all archaeologists recognize as factors affecting site location and function usually ignored in predictive modelling?" (Kohler, 1988:19). He gives the answer a few pages later: "Given the subtleties and especially the fluidity of the socio-political environment, is it any wonder that archaeologists have chosen to concentrate on those relatively stable, "distorting" factors of the natural environment for locational prediction?" (Kohler, 1988:21).

In essence, the situation has not changed since Kohler made these remarks. The present practice of predictive modelling is still very much environmentally deterministic. Cultural variables are not included in the models, resulting in predictions ultimately based on physical properties of the current landscape.

Practitioners of 'traditional' predictive modelling have mostly resisted the conclusion that their models will not be adequate because they lack the input of non-environmental data (e.g. Kvamme, 1997). It is not because they do not want to include non-environmental factors; the problem is that these variables are regarded as being too abstract and intangible for use in a predictive model. Such models, so the argument goes, will therefore not become any better by investing valuable research time in mapping cultural variables. Several publications have focused on this apparent impossibility to incorporate non-environmental variables in predictive modelling (Wheatley, 1996a; Stančič and Kvamme, 1999; and Lock 2000). As a consequence, very few studies are available where an attempt is made to improve the gain of a model by incorporating non-environmental factors. As a consequence, the effect of including cultural variables into predictive models can

at the moment not be assessed. The current situation is therefore characterized by a fundamental criticism of the environmental deterministic approach, coupled to a very poor insight into the potential of using cultural variables in predictive modelling.

Ultimately, the theoretical basis needed for the development of culturally based predictive models seems to be underdeveloped. It is evident that many models of prehistoric land use have been proposed for local case studies, but they are usually not generalized for application in a predictive modelling context, and often have never been tested in a rigorous way. A typical example of this is found in the theories regarding the location of Linear Band Ceramic settlements, in which a strong cultural component is supposed to be present (see Gaffney and van Leusen, 1995), yet no predictive model based on this assumption has ever been made.

In conclusion, it may be suspected that the lack of progress in incorporating cultural variables into predictive modelling has less to do with the variables themselves, than with the geographic and interpretative models needed to operationalize them for predictive modelling. Many applications that claim to be exponents of cognitive archaeology, often framed in post-processual rhetoric, rely on the same techniques that are used for old-fashioned, processual studies, up to the extent where they might even be called ‘cognitive deterministic’.

10.3. CULTURAL VARIABLES: WHAT ARE THEY?

It is important to realize that, when we are speaking of cultural variables, we can think of two ways of obtaining them. The first one is to consider them as measurable attributes of the archaeological sample that are not related to an environmental factor. So, instead of measuring for each individual site its soil type, elevation, distance from water and so on, we need to ask which properties of the site itself can be measured. These include properties like site location, size, functional type and period of occupation. These variables are clearly the expression of forms of social behaviour, although the interpretation of the specific behaviour involved may be subject to discussion. For ease of reference, these variables will be denominated cultural variables *sensu stricto*. In themselves, these variables are not extremely difficult to obtain, but the problems of analysing and interpreting archaeological site databases are manifold and must be addressed before these properties can actually be used for predictive modelling.

The second approach to defining cultural variables is to identify features of the landscape itself that can be interpreted as having cultural significance, such as sacred springs. These can be referred as to as cultural landscape variables, and are not necessarily excluded from ‘traditional’ predictive modelling, but often are not recognized as constituting a ‘cultural’ variable. It can, in fact, be argued that all environmental variables have a cultural component, even though the emphasis in traditional predictive modelling is usually on subsistence economy rather than symbolic meanings.

In order to make further use of cultural variables in predictive modelling, it is necessary to transform these variables into continuous variables: for each single variable a value should be available at any location within the study area. This is generally not a problem when using environmental data sets like soil maps or digital elevation models. Archaeological sites however are mostly represented as points, or in some cases as areas of a very limited extent. Similarly, landscape features that are considered to have cultural significance are in practice often also regarded as point-like, or at best linear in nature. A transformation is therefore necessary to use point-like or linear objects for predictive modelling. Two types of GIS techniques are currently available to perform this transformation: distance zonation and line-of sight analysis.

Distance zonation is customarily performed in environmental predictive modelling to obtain continuous variables from environmental features that are either linear (like rivers or coastlines) or point-like (springs).

In some cases, cost surfaces (also known as friction surfaces or effort models) are calculated by assigning a weight to landscape features according to their supposed accessibility. This technique is applicable to environmental as well as cultural variables.

Distance decay models are used less often, and are based on demographic and/or political-economic models borrowed from human geography (e.g. Renfrew and Level, 1979). These models are specifically relevant for cultural variables *sensu stricto*, as they make it possible to incorporate the notion of interdependence of settlements (see e.g. Favory *et al.*, 2003).

A number of studies have appeared in recent years using line-of-sight analysis as a technique for obtaining continuous cultural variables, amongst others in attempts to demonstrate the ritual and symbolic meaning of the placement of monuments such as long barrows (Wheatley, 1995; Gaffney *et al.*, 1995). However, this type of analysis is certainly not restricted to cultural variables.

A good example of the use of cultural variables *sensu stricto* and distance zonation is provided by Ridges (in press), who attempted to include the distance to rock art sites in a predictive model in NW Queensland (Australia) - and actually succeeded in improving the gain of the model. This success is probably due to the fact that the ritual sites used are fixed in space, and can be mapped with relative ease in the specific environmental situation. The rock art sites are typical examples of what Whitley (2000) refers to as 'fixed point attractors'. The precise moment of their creation may be unknown, but their position and symbolic meaning remain stable during a long period of time, making them long-term attractors for human activity[5].

In many other situations however, potential cultural variables are less stable, and cannot be mapped with ease. Examples of these include road networks, field systems, and the archaeological sites themselves, which all can have highly varying life-spans and may change in importance as attractors over time. In order to model the effects of long term land use development, it is necessary to use a technique that can deal with spatio-temporal variables, like dynamical systems modelling.

10.4. HOW TO PROCEED?

In order to remedy the current situation the following issues should be addressed:

- the identification of cultural variables that are significant for archaeological site location;
- the analysis of the utility of these variables for predictive modelling;
- the development and application of existing and new relevant modelling techniques; and
- the analysis of the performance of predictive models based on cultural variables compared to environmentally based models.

Following the recommendations in van Leusen *et al.* (2002), we suggest that four promising areas of research should be explored in order to improve on the current use of cultural variables in predictive modelling. These are:

[5] in the case of Aboriginal rock art sites, it might even be a combination of ecological and cultural factors, as the sites are supposed to have been used as 'markers', indicating the presence of natural resources

A systematic analysis of the archaeological records and their aggregation into culturally meaningful entities

It is necessary to analyse what information can be extracted from existing archaeological databases that can be used in the definition of cultural variables. The aggregation of the archaeological contents of find spots into meaningful archaeological entities is currently not standardized. A possible solution could be to design an expert system that can be used for the classification of find spots. Apart from defining meaningful archaeological entities, the aggregation of multiple find spots into single archaeological sites is an important issue where the utility of the archaeological database for predictive modelling is concerned. Thirdly, a tendency can be observed recently to combine multiple archaeological sites into ensembles, which effectively constitutes a step away from the site level and towards a regional, landscape-based concept of archaeological entities (see also Kuna, 2000).

The main question here is: what types of aggregates can we distinguish, and can these be used as cultural variables *sensu stricto*?

Analysis of the logistic position of settlements

It is anticipated that one of the most important cultural variables that can be used is the logistic position of the archaeological site itself. It has been shown by many researchers that the position of a settlement in a logistic network determines to a large degree its size and duration of occupation (e.g. Durand-Dastès *et al.*, 1998). The development of techniques to analyse the logistic position of settlements can be addressed by looking at recent work in human geography.

The continuity of the cultural landscape

The cultural landscape has a historical dimension that strongly influences its use and usability. The existing cultural landscape influences the positioning of new sites. Kuna (1998), for example, mentions the importance of remnants of past landscapes on settlement location choice. Bell *et al.* (2002) demonstrated how later settlement in their Central Italian study area avoids areas settled in an earlier phase but conforms to paths from that earlier phase. Techniques to perform the long-term diachronical analysis needed for this type of modelling have been developed (e.g. by the Archaeomedes project; van der Leeuw, 1998; Favory *et al.*, 2003).

Line-of-sight analysis

In hilly areas and with certain site types that have a strong visual component (like burial mounds or megalithic tombs) line-of-sight analysis may be a type of analysis suitable for predictive modelling (see van Leusen, 2002: chapters 6 and 16). The techniques for performing this type of analysis are well established.

It will be noticed that the four research topics mentioned here all focus on cultural variables *sensu stricto*. A thorough investigation of the use of cultural landscape variables would primarily involve the development of a decision rule framework that will incorporate the perception of the landscape into predictive modelling. In itself, this is an issue that merits attention, but the establishment of decision rules has always been at the heart of predictive modelling and is covered by a wide range of studies already. It would however be useful to start thinking about ways to model the perception of the landscape, as has been done by Whitley

(2000), who tried to model the attractivity of the landscape for specific (economic) activities of Native American hunter-gatherers (see also Whitley, in press).

10.5. CONCLUSIONS

In a recent article on the use and abuse of statistical methods in archaeological site location modelling Woodman and Woodward (2002) come to the following conclusion: "There has been much criticism of locational studies since they are often based largely on environmental criteria. However, before researchers attempt to incorporate the more intangible social, cognitive, political and aesthetic factors, it would be wise to employ the appropriate statistical techniques required to deal with the complexities which already exist in even the most basic tangible and quantifiable environmental criteria".

Although we do not deny that many statistical problems still exist with regard to predictive modelling, we see no apparent reason why they should receive prime importance in further developing predictive modelling. In fact, the three main issues of statistical methodology, the development of adequate archaeological (and non-archaeological) data sets and the incorporation of non-environmental factors into the models are closely connected, and cannot be tackled in isolation. The papers presented in van Leusen and Kamermans (in press) show that new approaches to predictive modelling are starting to emerge, like exploring the potential of Bayesian statistical methods, using high resolution data for predictive modelling, and looking for ways to better embed predictive models into archaeological heritage management practice, for example by developing risk assessment methods. There is no doubt still a lot to do, and in this respect we have to disagree with Wheatley (2003) who argues that too much money is going into predictive modelling studies. He may be right that funding for GIS-related archaeological projects is mainly going into predictive modelling, but compared to the amount of money spent on all forms of prospection and excavation, investments made in predictive modelling seem relatively modest. Apart from that, investments for a thorough, scientific analysis of predictive modelling have been few and discontinuous.

We hope to have demonstrated that incorporating cultural variables into predictive modelling can be done, even though it is impossible to present a comprehensive overview in these few pages. It is up to the scientific community and public institutions to decide if this line of research is worth investing in. However, if the three issues mentioned above (statistical improvements, quality of the archaeological data set and the development of non-environmentally based models) are not tackled in the years to come, predictive modelling will remain to be criticized as a tool that is of dubious scientific quality, and not even capable of providing clear answers on where to spend money for archaeological research.

REFERENCES

Bell, T., A. Wilson and A. Wickham, 2002. 'Tracking the Samnites: landscape and communications routes in the Sangro Valley, Italy'. *American Journal of Archaeology* 106 (2), pp. 169-186.

Durand-Dastès, F., F. Favory, J.-L. Fiches, H. Mathian, D. Pumain, C. Raynaud, L. Sanders and S. van der Leeuw, 1998. *Des oppida aux métropoles. Archéologues et géographes en vallée du Rhône.* Anthropos, Paris.

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

Favory, F., J.-L. Fiches and S. van der Leeuw, 2003. Archéologie et systèmes socio-environnementaux. Etudes multiscalaires sur la vallée du Rhône dans le programme ARCHAEOMEDES. CRA-Monographies. CNRS Editions, Paris.

Gaffney, V. and M. van Leusen, 1995. 'Postscript - GIS, environmental determinism and archaeology: a parallel text', in: Lock, G. and Z. Stančič (eds.), *Archaeology and Geographical Information Systems: A European Perspective*. Taylor & Francis, London, pp. 367-382.

Gaffney, V., Z. Stančič and H. Watson, 1995. 'The impact of GIS on archaeology: a personal perspective', in: G. Lock and Z. Stančič (eds.), *Archaeology and Geographical Information Systems: A European Perspective*. Taylor & Francis, London, pp. 211-229.

Kohler, T.A., 1988, 'Predictive locational modelling: history and current practice', in: Judge, W.L. and L. Sebastian (eds.), *Quantifying the Present and Predicting the Past: Theory, Method and Application of Archaeological Predictive Modeling*. US Bureau of Land Management, Denver, pp. 19-59.

Kuna, M., 1998. 'The Memory of Landscapes', in: Neustupný, E. (ed.), *Space in Prehistoric Bohemia*. Academy of Science, Prague, pp. 77-83.

Kuna, M., 2000. 'Surface Artefact Studies in the Czech Republic', in: Bintliff, J., M. Kuna and N. Venclová (eds.), *The future of surface artefact survey in Europe.* Sheffield Archaeological Monographs 13. Sheffield Academic Press, Sheffield, pp. 29-44.

Kvamme, K.L., 1997. 'Ranters Corner: bringing the camps together: GIS and ED'. *Archaeological Computing Newsletter* 47: 1-5.

Leeuw, S.E. van der (ed.), 1998. *The Archaeomedes Project: Understanding the natural and anthropogenic causes of soil degradation and desertification in the Mediterranean Basin. Research Results.* Office for Official Publications of the European Communities, Luxembourg.

Leusen, M. van, 2002. *Pattern to Process. Methodological investigations into the formation and interpretation of spatial patterns in archaeological landscapes.* Rijksuniversiteit Groningen, Groningen. PhD thesis.

Leusen, 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 for the BBO research program.* Interim report. Rijksuniversiteit Groningen, Groningen.

Leusen, M. van, and H. Kamermans (eds.), in press. *Predictive Modelling for Archaeological Heritage Management: A Research Agenda*. Nederlandse Archeologische Rapporten. Rijksdienst voor het Oudheidkundig Bodemonderzoek, Amersfoort.

Lock, G. (ed.), 2000. *Beyond the Map. Archaeology and Spatial Technologies*. NATO Science Series, Series A: Life Sciences - Vol. 321. IOS Press, Amsterdam

Renfrew, C., and E.V. Level, 1979. 'Exploring dominance: predicting polities from centers', in: Renfrew, C. and K.L. Cooke (eds.), *Transformations: Mathematical approaches to culture change.* Academic Press, New York, pp. 145-167.

Ridges, M., in press. 'Understanding H-G behavioural variability using models of material culture: An example from Australia', in: Mehrer, M. and K. Wescott (eds.), *GIS and Archaeological Site Location Modeling.* CRC Press, Boca Raton, Florida USA.

Stančič, Z. and K.L. Kvamme, 1999. 'Settlement Pattern Modelling through Boolean Overlays of Social and Environmental Variables', in: Barceló, J.A., I.Briz and A. Vila (eds.), *New Techniques for Old Times -CAA98. Computer Applications and Quantitative Methods in Archaeology*. BAR International Series 757. Archaeopress, Oxford, pp. 231-237.

Verhagen, P., in press. 'Prospection strategies and archaeological predictive modelling', in: Leusen, M. van and H. Kamermans (eds.), *Predictive Modelling for Archaeological Heritage Management: A Research Agenda.* NAR, ROB, Amersfoort.

Wheatley, D. 1993, 'Going over old ground: GIS, archaeological theory and the act of perception', in: Andresen, J., T. Madsen, and I. Scollar (eds.), *Computing the Past: Computer Applications and Quantitative Methods in Archaeology - CAA 92.* Aarhus. 133-38.

Wheatley, D., 1995. 'Cumulative viewshed analysis: a GIS-based method for investigating intervisibility, and its archaeological applications', in: Lock, G. and Z. Stančič (eds.), *Archaeology and Geographical Information Systems: A European Perspective*. Taylor & Francis, London, pp. 171-185.

Wheatley, D., 1996a. 'Between the lines: the role of GIS-based predictive modelling in the interpretation of extensive survey data', in: Kamermans, H. and K. Fennema (eds.), *Interfacing the Past. Computer applications and quantitative methods in Archaeology CAA95*. Analecta Praehistorica Leidensia 28: 275-292.

Wheatley, D., 1996b. 'The use of GIS to understand regional variation in Neolithic Wessex', in: Maschner, H.D.G. (ed.), *New methods, old problems: Geographic Information Systems in modern archaeological research.* Occasional Paper No. 23. Center for Archaeological Investigations, Southern Illinois University, Carbondale (IL), pp. 75-103.

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., 2000. *Dynamical Systems Modeling in Archaeology: A GIS Approach to Site Selection Processes in the Greater Yellowstone Region.* Unpublished PhD thesis. Department of Anthropology, University of Pittsburgh, Pittsburgh (PA).

Whitley, T.G., in press. 'A Brief Outline of Causality-Based Cognitive Archaeological Probabilistic Modeling', in: Leusen, M. van and H. Kamermans (eds.), *Predictive Modelling for Archaeological Heritage Management: A Research Agenda.* Nederlandse Archeologische Rapporten. Rijksdienst voor het Oudheidkundig Bodemonderzoek, Amersfoort.

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

POSTSCRIPT TO CHAPTER 10

Part of this paper was originally written as a grant proposal for the second phase of the BBO programme. Unfortunately, the research suggested in the paper was not funded, and we have made no further attempts to find other sources of funding. The type of research advocated in this paper is not a priority in Dutch archaeology, and perhaps not even in international archaeology. It is difficult to say why, as the reviews of the grant proposal by external experts were positive, and its scientific and societal relevance was considered high by the review committee. The main objection brought forward against the proposal was the fact that the proposed research could not guarantee a successful outcome, and underestimated the complexity of the matter, so perhaps even needed more funding than was asked for.

However, there is a strong case for doing this type of research, as is explained in the second section of the paper. The post-processual critique of archaeological predictive modelling is mainly based on the conviction that ecological factors cannot offer a full explanation, and therefore not a valid prediction, of site location preferences. This ignores the fact that environmentally based predictive modelling, and related 'environmental' methods like site catchment analysis, have been quite successful, provided they use data sets of sufficient quality. But obviously, any predictive model will have a 'residual' of sites that do not fit the (environmental) explanatory framework applied, and these are the sites that should be analysed for other factors, including socio-cultural ones. Post-processual theorists however have largely remained silent when it comes to finding a way of integrating socio-cultural factors into predictive modelling. While we are certainly dealing with a complex matter, it seems that earlier attempts to deal with it have focused too much on matters that are truly intangible, like the perception of the landscape in the minds of prehistoric people. Our approach therefore was a more pragmatic one: given the available 'cultural variables', can we try to develop predictive models that perhaps will not cover *all* aspects of site location theory, but that will at least contribute to a better prediction? Unfortunately, we will have no opportunity to find out, at least not in the near future.

EPILOGUE WHITHER ARCHAEOLOGICAL PREDICTIVE MODELLING?

The BBO Predictive Modelling project has now come to an end, and if anything has become clear, it is that predictive modelling is an issue that is far from 'solved'. The project has been successful in defining the problem areas, and has contributed to a better understanding of why predictive modelling is so much debated. It also has experimented with some exciting new approaches, and has established fruitful connections with practitioners of predictive modelling outside the Netherlands. However, the project has up to now failed to realize any significant changes in predictive modelling practice in the Netherlands. Part of this is due to the scientific outlook of the programme. It is of primary importance that the results of scientific research are published in English and presented at international conferences. However, a national impact still has to materialize, and this is partly the consequence of a lack of discussion in Dutch forums.

I also have to conclude that academic and public archaeology are still opposed when it comes to predictive modelling. Whereas commercial parties and the ROB continue to make and use predictive models like they have done over the past fifteen years, academic criticism has not stopped and will not stop unless some fundamental improvements are made to current practice, especially where the IKAW is concerned. This is not easy: the methods explored in the project and by other researchers require a lot of work, both at the fundamental level of applying them in the right way, as well as in the collection and screening of archaeological data; the development of new site location theories; and the formulation of quality demands.

The perspective of predictive modelling in the Netherlands can be sketched in three scenarios for the near future. The first one is a scenario of 'business as usual' and assumes that predictive models and maps will continue to be made and used in Dutch archaeological heritage management like they have been over the past decade. 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. They are used for the designation of zones of archaeological interest in town and country plans; they play an important role in environmental impact assessments; and they are now even published free of charge for the general audience on the new internet portal for cultural-historical information (www.kich.nl). 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: the province of Limburg recently made public that it wants to change its policy to do survey in 70% of the province (the area designated as high and medium probability) - the main argument being that about 45% of the archaeological survey projects carried out in the province have not resulted in the discovery of any archaeology at all (*Dagblad De Limburger*, 20 September 2005). This new policy is supposed to result in an overall reduction of 60% of the total amount of money spent on archaeology. So, predictive maps are in fact more than effective in protecting the archaeological heritage: they over-protect. And 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. Alderman B. Pauwels of the municipality of Hulst (province of Zeeland) put it as follows:

'We're talking about a large area. Under the current circumstances, in every project, no matter its size, archaeological research must be done. This is a complex matter and it takes time and money (…) By using a predictive map, we can estimate where to expect valuable archaeological remains in the soil. It economizes on preliminary research. At the moment, we don't know what to expect.' (excerpt translated from the *Provinciale Zeeuwse Courant*, 22 September 2004).

Current efforts in predictive model refinement are therefore geared towards increasing the resolution of the mapping, and a better definition of the zones of no interest, both as a means to limit the zones of high probability to the absolute minimum. This development however demonstrates the risk of continuing business as usual: when predictive maps are primarily based on expert judgement, or on bad data sets, how can one tell if the province of Limburg really needs to protect 70% of its territory? What criteria are used to decide how large the area of high probability should be? 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. It is a scenario that has been defended by Wheatley (2003), and 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 (Altschul *et al.*, 2004). 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, like 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. Moreover, the idea that full survey coverage will detect all archaeological sites is naïve and wrong (Altschul *et al.*, 2004; chapter 6). So, while attractive from the archaeologists' point of view, full survey is not a practical alternative to predictive modelling in the real world. The question then is: is selection by means of predictive maps better than doing the desk-based assessments that are common practice in France and the United Kingdom? The answer to this question is affirmative: it is better to have a map that draws the attention to potential archaeology, than a map that only indicates the location of known sites – even if this predictive map does not provide a 100% correct prediction. 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. Clearly, we still need to prove that models based on statistical estimates, that also provide a measure of the uncertainty involved, will actually do a better job in archaeological heritage management than the currently available expert judgment and relative site density maps. There are, however, 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. At the moment, expert judgment is determining 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.

The main objections against this innovation of predictive modelling are financial and psychological. Financial, because making predictive models this way implies a thorough analysis of the archaeological data, and the use of rather complex statistical techniques, which will make model building more time consuming than expert judgment classification. And psychological, as statistical models are often seen as ‘black box’ models; the suspicious attitude to statistical methods in archaeology is deeply rooted, as it is sometimes felt that statistics ‘can be used to prove anything’, and as a consequence can never be trusted. While this is a misunderstanding of the nature of statistical methods and the results they produce, we will still need a lot of publicity and good results to convince the archaeological community. So, there is a lot of work to do ….

REFERENCES

Altschul, J.F., L. Sebastian and K. Heidelberg, 2004. *Predictive Modeling in the Military. Similar Goals, Divergent Paths.* Preservation Research Series 1. SRI Foundation, Rio Rancho (NM). http://www.srifoundation.org/pdf/FINALLEG.pdf, accessed on 24-11-2005.

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

SAMENVATTING

Dit proefschrift, getiteld "*Case studies* van archeologische voorspellingsmodellen", heeft als thema het verbeteren van de modelleringstechnieken en toetsingsmethoden die gebruikt kunnen worden bij het maken van archeologische verwachtingskaarten. Deze kaarten worden al sinds de jaren zeventig van de 20e eeuw gemaakt in de Verenigde Staten, en worden vanaf circa 1990 in de Nederlandse archeologische monumentenzorg gebruikt. Zij doen een uitspraak over de mogelijke aanwezigheid van archeologische resten op plaatsen waar nog geen archeologisch onderzoek is gedaan. Deze voorspellingen zijn gebaseerd op een analyse van de ligging van reeds bekende archeologische vindplaatsen ten opzichte van factoren als bodemgesteldheid of de nabijheid van stromend water, en/of op aannames omtrent het belang van deze factoren voor locatiekeuze. Tegenwoordig is het in Nederland gebruikelijk om met behulp van deze kaarten te besluiten of archeologisch vooronderzoek noodzakelijk is. Als uit de voorspelling blijkt dat er een zeer kleine kans is op de aanwezigheid van archeologische resten in de bodem, dan zal dit vooronderzoek vaak achterwege blijven. Daarnaast worden archeologische verwachtingskaarten ook gebruikt om afwegingen te maken bij milieu effect rapportages: met behulp van een verwachtingskaart kan bijvoorbeeld het tracé worden bepaald waardoor de archeologie het minst verstoord zal worden.

Ondanks deze algemene acceptatie en het gebruiksgemak van archeologische verwachtingskaarten, zijn zij verre van omstreden in de archeologische wereld. Dit heeft te maken met de veronderstelde kwaliteit van de voorspellingen. In de praktijk blijkt dat de statistische en conceptuele modellen die gebruikt worden om de voorspellingen te doen, vaak gebaseerd zijn op onvolledige gegevensbestanden en op gebrekkige theorieën omtrent de factoren die bepalen waarom zich ergens archeologische vindplaatsen bevinden.

Het proefschrift bestaat uit een bundeling van artikelen die zijn geschreven in een tijdsperiode van acht jaar, van 1997 tot en met 2005. De auteur werkte in deze periode bij RAAP Archeologisch Adviesbureau. RAAP heeft aan dit proefschrift bijgedragen door het beschikbaar stellen van onderzoekstijd en van fondsen om wetenschappelijke congressen te bezoeken. Daarnaast zijn de artikelen mede mogelijk gemaakt dankzij het geld van verschillende opdrachtgevers, waarbij met name de bijdrage van het onderzoeksprogramma *Bodemarchief in Behoud en Ontwikkeling* van NWO moet worden genoemd, en die van het onderzoeksprogramma *Archaeomedes*, gefinancierd door de Europese Unie. Omdat het gaat om diverse projecten en bijdragen, die over een lange periode zijn uitgevoerd, zijn de artikelen gerangschikt in drie thematische blokken, en voorzien van een uitgebreid inleidend hoofdstuk over de achtergrond en geschiedenis van archeologische voorspellingsmodellen. Het proefschrift bestaat uit drie thematische blokken, die vooraf worden gegaan door een inleidend hoofdstuk. Deze zullen in deze samenvatting kort worden besproken.

HOOFSTUK 1: EEN BEKNOPTE GESCHIEDENIS VAN ARCHEOLOGISCHE VOORSPELLINGS-MODELLEN

In dit hoofdstuk wordt uitgebreid ingegaan op de geschiedenis en achtergrond van archeologische voorspellingsmodellen. Aan de orde komen:

- de redenen voor het maken en gebruiken van archeologische verwachtingskaarten;
- het onderscheid tussen inductief en deductief modelleren;
- de opkomst van de archeologische monumentenzorg;

- de invloed van *New Archaeology* met haar nadruk op kwantitatieve methoden;
- de verbreiding van het gebruik van Geografische Informatie Systemen in de archeologie;
- en de post-processuele kritiek op archeologische voorspellingsmodellen.

Ook wordt duidelijk gemaakt dat er op het gebied van archeologische verwachtingskaarten een patstelling is ontstaan tussen de (kritische) academische wereld enerzijds en de archeologische monumentenzorg anderzijds, waar deze kaarten op grote schaal worden gemaakt en gebruikt. De belangrijkste bezwaren die tegen het maken en gebruiken van archeologische verwachtingskaarten worden aangevoerd zijn:

- het gebruik van onvolledige archeologische gegevensbestanden;
- een selectieve keuze van de veronderstelde factoren die invloed hebben op locatiekeuze, die vooral gebaseerd is op de beschikbaarheid van goedkope digitale datasets;
- onderschatting van de invloed van sociaal-culturele factoren op locatiekeuze;
- onderschatting van de dynamiek van het landschap.

Hier staat tegenover dat er een duidelijke maatschappelijke vraag is naar voorspellingen die met enige mate van zekerheid kunnen aangeven waar zich archeologische resten kunnen bevinden. De kwestie is dus niet zozeer of het wel mogelijk is om archeologische verwachtingskaarten te maken, maar of deze binnen de beperkingen van de gebruikte gegevensbestanden en theoretische uitgangspunten tot een acceptabele voorspelling leiden. Tot op heden zijn er echter nog geen oplossingen aangedragen die zowel praktisch uitvoerbaar en betaalbaar zijn, als voldoen aan alle eisen van wetenschappelijke kwaliteit van de voorspellingskaarten. Vervolgens wordt nader ingegaan op de achtergrond van het onderzoeksproject *Strategic research into, and development of best practice for, predictive modelling on behalf of Dutch cultural resource management*, dat vanaf 2002 heeft geprobeerd om deze eisen nader tot elkaar te brengen. Hoewel de resultaten van het project aanleiding geven om te veronderstellen dat dit wel degelijk mogelijk is, ontbreekt het tot op heden aan praktische uitvoering van de aanbevelingen die in het kader van dit project worden gedaan.

DEEL 1: PRAKTISCHE TOEPASSINGEN

In dit deel zijn drie artikelen samengebracht, die elk een *case study* presenteren met verschillende invalshoeken voor het maken van archeologische voorspellingsmodellen.

HOOFSTUK 2

In hoofdstuk 2 wordt een voorspellingsmodel besproken dat is gemaakt voor het Argonne-gebied in Noordoost-Frankrijk. Dit model had als doel om de locatie te voorspellen van pottenbakkersovens, die in de Romeinse tijd werden gebruikt voor de massaproductie van aardewerk dat naar grote delen van het noordwesten van het Romeinse rijk werd geëxporteerd. Voorafgaand aan het project was zeer weinig informatie beschikbaar, en daarom werd gekozen voor het uitvoeren van archeologische prospectie (vooral veldkartering) op basis van een voorspellingsmodel. Dit bleek een gelukkige keuze te zijn: het voorspellingsmodel kon worden gebruikt om de zones aan te wijzen die nog onvoldoende gekarteerd waren. Tijdens het project werd steeds duidelijker dat de locatie van de ovens vooral gerelateerd was aan de aanwezigheid van geschikte kleivoorkomens op niet al te grote afstand van stromend water. Deze factoren

konden met behulp van geologische en topografische kaarten van het gebied goed in kaart worden gebracht. Kleine onvolkomenheden in het model zijn vooral het gevolg van onvolledige informatie in het bronnenmateriaal. Daarnaast bleek een klein gedeelte van de ovens primair in de buurt van transportroutes te liggen. Doordat het modelleren en de prospectie nauw op elkaar aansloten, was het mogelijk om uiteindelijk een model te maken met een hoge voorspellende waarde. De keerzijde was dat deze manier van werken tijdrovend is, en tot hoge prospectiekosten leidt.

HOOFSTUK 3

Hoofdstuk 3 beschrijft hoe in een ander gebied in Frankrijk, de Tricastin-Valdaine regio aan de oostoever van de Rhône, is geprobeerd om aan te tonen hoezeer oppervlaktekartering een vertekend beeld kan geven van de locatie en de hoeveelheid vindplaatsen in alluviale gebieden. Hoewel dit in Nederland al langer bekend was, was dit voor de Franse situatie niet het geval, en voorzover bekend is het ook de enige keer dat geprobeerd is om dit effect te kwantificeren. Met behulp van gedetailleerde archeologische vindplaatsgegevens is op basis van geomorfologische en bodemkundige kaarten een analyse uitgevoerd van de ligging van de bekende vindplaatsen. Daarbij werd telkens een aparte analyse uitgevoerd voor de vindplaatsen die aan de oppervlakte lagen, en de begraven vindplaatsen. Ook werd gekeken naar de vertekening die optreedt als de analyse wordt uitgevoerd zonder dat bekend is welke delen van het gebied zijn gekarteerd, ten opzichte van een analyse die wel rekening houdt met de mate waarin de verschillende zones in het gebied zijn gekarteerd. In de alluviale zones bleken veel meer vindplaatsen voor te komen dan algemeen werd verondersteld, en dit kon ook kwantitatief worden onderbouwd. Verder is geprobeerd om een ruwe schatting te maken van het aantal vindplaatsen dat nog in het gebied aanwezig zou kunnen zijn. Hieruit kwam naar voren dat in het bestudeerde gebied naar alle waarschijnlijkheid nog ongeveer 8625 onontdekte archeologische vindplaatsen liggen.

HOOFSTUK 4

Dit hoofdstuk richt zich op het probleem van het gebruik van het oordeel van experts voor het maken van voorspellingen. Omdat in lang niet alle gevallen gebruik kan worden gemaakt van gecontroleerde steekproeven van voldoende omvang, is het in Nederland steeds meer de gewoonte geworden om voorspellingsmodellen te maken op basis van de veronderstellingen van deskundigen over het belang van verschillende locatiefactoren. Dit wordt ook wel aangeduid als *expert judgement*. Deze methode heeft echter als groot nadeel dat zij nauwelijks controleerbaar is, en bovendien slecht te combineren is met kwantitatieve technieken. De hypothese die in dit hoofdstuk wordt uitgewerkt, is dat in deze gevallen gebruik kan worden gemaakt van Bayesiaanse statistische methoden, die in staat zijn om subjectieve oordelen te combineren met gecontroleerde steekproefgegevens. Hiervoor is wel noodzakelijk dat de experts hun oordeel in statistische termen kunnen weergeven. Met andere woorden, zij moeten getalsmatig kunnen aangeven hoe zij denken over het belang van verschillende locatiefactoren, en hoe zeker zij daarvan zijn. Deze inschatting wordt aangeduid als de *a priori* kansverdeling. Door het later toevoegen van gecontroleerde steekproefgegevens valt de aanvankelijke schatting zowel te toetsen als te verfijnen. Dit staat bekend als het vaststellen van de *a posteriori* kansverdeling. In de in dit hoofdstuk gepresenteerde *case study* is de a priori kansverdeling bepaald door een drietal experts een voorspelling te laten doen voor het grondgebied van de gemeente Ede met behulp van technieken uit de multi-criteria analyse. Dit bleek een effectieve manier om de experts een a priori

kansverdeling op te laten stellen, maar kon ook gebruikt worden om te laten zien in hoeverre de experts elkaar tegenspraken. Door gebrek aan gecontroleerde steekproefgegevens in het studiegebied is de fase van het vaststellen van a posteriori kansverdeling blijven steken in een demonstratie van de toepassingsmogelijkheden. Verder bleek dat het op de juiste wijze toepassen van Bayesiaanse statistiek complex is, en nog nadere studie behoeft. Het grote voordeel is wel dat op grond van algemene uitspraken van experts een statistische schatting kan worden gegeven met een betrouwbaarheidsmarge. Het is juist deze betrouwbaarheidsmarge die bij de huidige voorspellingsmodellen vrijwel altijd ontbreekt.

DEEL II: ARCHEOLOGISCHE PROSPECTIE, STEEKPROEVEN EN VOORSPELLINGSMODELLEN

In dit deel worden de resultaten gepresenteerd van twee onderzoeken, die betrekking hebben op het nemen van steekproeven met behulp van archeologische prospectie, en de consequenties hiervan voor het maken en toetsen van archeologische voorspellingsmodellen.

HOOFSTUK 5

In hoofdstuk 5 wordt een korte inleiding gegeven op de statistische achtergrond van het vraagstuk van de optimale prospectiemethode bij archeologische booronderzoek. Dit artikel is een uitvloeisel van het door Senter gefinancierde onderzoek naar de effectiviteit van booronderzoek voor archeologische prospectie, dat heeft geresulteerd in een Nederlandstalige publicatie[1]. De kern van het hoofdstuk wordt gevormd door de boodschap dat booronderzoek vrijwel nooit tot een volledige opsporing van archeologische vindplaatsen kan leiden. Slechts vindplaatstypen die zich kenmerken door een relatief grote omvang en een sterke concentratie van archeologische indicatoren zullen altijd worden opgespoord met booronderzoek, voor andere typen vindplaatsen ligt die kans veel lager. Daarom is het van het grootste belang dat voorafgaand aan prospectie wordt gedefinieerd naar welk type vindplaats er gezocht gaat worden, in termen van omvang en hoeveelheid verwachte archeologische indicatoren.

HOOFSTUK 6

In hoofdstuk 6 wordt deze constatering verder uitgebouwd door ook te kijken naar alternatieven voor booronderzoek. Ook oppervlaktekartering en kartering door middel van proefsleuven hebben hun beperkingen, en het kiezen van de juiste prospectiemethode is daarmee altijd een kwestie van afwegen tussen effectiviteit en kosten. Opvallend is echter dat deze afweging in de praktijk nooit op grond van kansberekening wordt gemaakt. Zelfs algemene richtlijnen als het percentage dekking door middel van proefsleuven zijn in de praktijk bepaald door jarenlange gewoonte, en niet door een inschatting van de kans op het aantreffen van een bepaald type vindplaats. Dit heeft op zijn beurt weer consequenties voor de kwaliteit van de archeologische vindplaatsenbestanden die gebruikt worden voor het maken van voorspellingsmodellen. Het succes van archeologische prospectie staat of valt met de gebruikte prospectiemethode. Elke methode kent blinde hoeken, en pas door het analyseren van deze blinde hoeken kan duidelijk worden in hoeverre een prospectieonderzoek

[1] Tol, A., P. Verhagen, A. Borsboom & M. Verbruggen, 2004. *Prospectief boren. Een studie naar de betrouwbaarheid en toepasbaarheid van booronderzoek in de prospectiearcheologie*. RAAP-rapport 1000. RAAP Archeologische Adviesbureau, Amsterdam.

een volledig beeld van de aanwezigheid van archeologische vindplaatsen geeft. In de praktijk vindt deze analyse zelden of nooit plaats.

HOOFSTUK 7

Dit hoofdstuk is geschreven als detailstudie in het kader van het project *Strategic research into, and development of best practice for, predictive modelling on behalf of Dutch cultural resource management.* Het gaat in op de vraag hoe de voorspellingen die op een archeologische verwachtingskaart staan aangegeven kunnen worden getoetst. In de praktijk blijkt wel af en toe toetsing van de voorspellingen plaats te vinden, maar deze is niet onderbouwd met statistische argumenten en wordt zelden of nooit teruggekoppeld naar het voorspellingsmodel. Bovendien is onduidelijk wat de omvang van de toetsing zou moeten zijn om tot een aanvaardbare betrouwbaarheid van de voorspelling te kunnen komen. In deze materie speelt ook mee dat afweging op grond van kansberekening in de dagelijkse archeologische praktijk geen rol van betekenis speelt. Toch wordt met behulp van de bestaande verwachtingskaarten wel een beslissing genomen over het wel of niet uitvoeren van prospectie. Daarbij wordt de vraag naar de betrouwbaarheid soms wel gesteld, maar zijn er geen instrumenten beschikbaar om deze te kwantificeren. In dit hoofdstuk worden drie aspecten van toetsing uitgebreid besproken.

Het meten van de kwaliteit van voorspellingen

Een goed voorspellingsmodel leidt tot een verwachtingskaart waarin zoveel mogelijk archeologische vindplaatsen zich bevinden in een zo klein mogelijke zone van ‘hoge archeologische verwachting’. Het eerste aspect wordt ook wel aangeduid als *accuracy* (nauwkeurigheid), en het tweede aspect als *precision* (precisie). In het verleden zijn meerdere methoden ontwikkeld om deze twee aspecten te meten en vergelijkbaar te maken. Uit het onderzoek blijkt dat van de ontwikkelde maten de veel gebruikte *gain* hiervoor het meest geschikt is, maar dat deze desondanks niet in staat is om alle situaties onderling te vergelijken. *Gain* gaat er namelijk van uit dat het belang van nauwkeurigheid (het aantal vindplaatsen in de zone van hoge verwachting) even groot is als het belang van precisie (de omvang van de zone van hoge verwachting). In de praktijk is er echter een spanningsveld tussen de archeologische monumentenzorg, die gebaat is bij een zo hoog mogelijke nauwkeurigheid, en de economische en politieke realiteit, die vraagt om een zo precies mogelijk voorspellingsmodel. Vanuit archeologisch perspectief gezien is het daarom aan te bevelen om eerst een minimale nauwkeurigheid als norm voor een verwachtingskaart op te stellen, en daarbinnen te proberen een zo hoog mogelijke precisie te bereiken. Op die manier zijn voorspellingsmodellen ook beter onderling vergelijkbaar, en zijn zij niet onderhandelbaar in de zin dat archeologische vindplaatsen kunnen worden ‘ingeleverd’ voor een hogere precisie.

Kwaliteitsverbetering zonder onafhankelijke toetsing

Het tweede aspect dat nader wordt uitgewerkt in hoofdstuk 7 is de mogelijkheid om de voorspellende waarde van verwachtingskaarten te vergroten zonder nieuwe, onafhankelijke gegevens te verzamelen. De hiervoor beschikbare methoden maken gebruik maken van een statistische techniek die bekend staat als *resampling*. Hierbij wordt uit de beschikbare gegevens steeds een nieuwe steekproef genomen. Deze kunstmatige gecreëerde steekproeven kunnen dan worden vergeleken om een schatting van de

betrouwbaarheid van de voorspellingen te produceren. Deze methoden zijn in de archeologische literatuur sterk bekritiseerd, juist omdat zij geen gebruik maken van onafhankelijke gegevens. Hoewel het inderdaad zo is dat formele statistische toetsing gebruik dient te maken van een onafhankelijk gegevensbestand, is het niet zo dat deze *resampling*-methoden daarom onbruikbaar zijn. Sterker nog, in de statistiek zijn deze methoden de laatste jaren sterk in opkomst, omdat zij binnen de beperkingen van de gebruikte gegevens de betrouwbaarheid van voorspellingen sterk kunnen verbeteren. Het is daarom aan te bevelen om deze methoden standaard in te zetten bij het bepalen van de kwaliteit van voorspellingsmodellen.

Toetsen met onafhankelijke gegevensbestanden

Als laatste onderwerp wordt in hoofdstuk 7 ingegaan op het gebruikmaken van nieuwe, onafhankelijke gegevens voor toetsing. Hoe groot moet de omvang zijn van de te nemen steekproef om voorspellingen met voldoende betrouwbaarheid te krijgen? Twee complicerende factoren zijn hierbij van belang. Ten eerste worden de huidige verwachtingskaarten niet gepresenteerd met betrouwbaarheidsmarges. De toepasbaarheid van statistische methoden wordt daarmee sterk beperkt, omdat niet duidelijk is wat de werkelijke en wat de gewenste betrouwbaarheid van de voorspelling is. Daarnaast is het zo dat zelfs indien een gewenste betrouwbaarheid bekend is, het moeilijk is om vooraf te voorspellen hoe groot het gebied is dat geprospecteerd moet worden om een gegevensbestand van voldoende omvang voor toetsing te verkrijgen. Alle statistische methoden om te bepalen wat de benodigde steekproefgrootte voor toetsing moet zijn gaan uit van absolute schattingen, gebruik makend van een gecontroleerde steekproef. Aan geen van beide voorwaarden wordt voldaan bij de huidige wijze van het maken van archeologische voorspellingsmodellen.

Om tot een goede schatting van de betrouwbaarheidsmarge en een bijbehorende steekproefgrootte voor toetsing te kunnen komen, moeten er dus voorspellingsmodellen worden gemaakt die schattingen maken van absolute aantallen vindplaatsen op basis van een op vertekeningen gecontroleerd en gecorrigeerd gegevensbestand. Een verkenning van de onderzoeksgegevens in ARCHIS leert dat dergelijke gegevens misschien wel te verkrijgen zijn, maar niet op basis van de in ARCHIS geregistreerde kenmerken. Aspecten als de grootte van het onderzoeksgebied, het aantal gezette boringen, of de diepte tot waarop onderzoek is verricht, worden namelijk niet systematisch geregistreerd. Het vergt daarom extra inspanning om de benodigde gegevens te verzamelen en te controleren.

DEEL III: ALTERNATIEVE MANIEREN OM VOORSPELLINGSMODELLEN TE MAKEN

In het derde deel van dit proefschrift zijn drie artikelen gegroepeerd die kijken naar alternatieve manieren om voorspellingsmodellen te maken.

HOOFSTUK 8

In hoofdstuk 8 wordt een *case study* gepresenteerd waarin wordt geprobeerd om de potentiële agrarische productiezones te reconstrueren in het Vera-bekken in Zuid-Oost Spanje. De vraagstelling die hieraan ten grondslag ligt richt zich vooral op landdegradatie op de lange termijn, en de modellering is daarom ook uitgevoerd voor alle relevante archeologische perioden, vanaf het Neolithicum tot en met de periode van de Arabische overheersing. Op grond van schattingen van de bevolkingsomvang voor elke archeologische site, en hypothesen over het voedselgebruik, de verbouwde gewassen, de productiviteit van de verschillende landschappelijke zones in het gebied, en de toegankelijkheid van het gebied, is voor elke bekende archeologische site een maximale omvang bepaald waarbinnen landbouw zou hebben kunnen plaatsvinden. Daarbij is ook gekeken naar het effect van irrigatiesystemen op de productiviteit. De modellering ondersteunde de hypothese dat er in de zgn. El Argar-periode (Midden-Bronstijd) problemen kunnen zijn ontstaan met de voedselvoorziening. Om de geschatte bevolkingsomvang te kunnen voeden, moesten volgens het model gebieden in cultivatie worden genomen die daarvoor ongeschikt waren, wat bovendien geleid kan hebben tot ontbossing en daardoor een toename van erosie. Voor de Romeinse tijd bleek duidelijk dat het gebied niet in staat kan zijn geweest om surplusproductie te genereren zonder de inzet van irrigatiesystemen. Toch is uit historische bronnen bekend is dat het gebied in die tijd landbouwproducten exporteerde. Opvallend is verder dat in de Arabische periode dankzij verfijnde irrigatiesystemen het benodigde areaal aan landbouwgrond relatief klein was.

Hoewel er in deze studie dus geen sprake was van een voorspellingsmodel zoals in deel I, biedt deze aanpak wel de mogelijkheid om verschillende scenario's door te rekenen, en deze te vergelijken met de archeologische realiteit. Eén opvallend resultaat van de modellering was het geringe voorspelde landgebruik in de omgeving van het stadje Turre. Daar zijn geen archeologische vindplaatsen bekend, terwijl volgens het model de omgeving uitstekend geschikt is voor landbouw. Mogelijk is Turre dus gebouwd op eerdere nederzettingen, die daardoor nog niet aan het licht zijn gekomen door archeologische veldkartering.

HOOFSTUK 9

In hoofdstuk 9 wordt globaal ingegaan op de gebruiksmogelijkheden van landevaluatie voor het opstellen van landgebruiksmodellen zoals die in hoofdstuk 8. Landevaluatie is een methode die gebruik maakt van meetbare kenmerken van de bodem, zoals vochtvasthoudend vermogen en vruchtbaarheid, om een inschatting te maken van de aantrekkelijkheid van deze bodems voor landbouw. In de archeologie is deze techniek weinig toegepast, hoewel het op zichzelf een nuttige methode kan zijn voor het voorspellen van landbouwkundig potentieel, en daarmee een bijdrage kan leveren aan het voorspellen van de mogelijke locatie voor archeologische vindplaatsen. In de praktijk is het opstellen van een prehistorische landevaluatie echter geen gemakkelijke opgave. Met behulp van een aantal voorbeelden wordt duidelijk gemaakt dat de benodigde

informatie voor landevaluatie vaak moeilijk te verkrijgen is voor de situatie in het verleden. Veranderingen in de waterhuishouding en erosie kunnen bijvoorbeeld het landbouwkundig potentieel sterk beïnvloeden. Deze zullen dus moeten worden gereconstrueerd om een prehistorische landevaluatie te kunnen uitvoeren. Daarnaast is niet gezegd dat de criteria die een moderne landevaluatie aanlegt ten aanzien van de productiviteit ook dezelfde zijn die in het verleden werden gehanteerd. Uit Romeins bronnenmateriaal blijkt bijvoorbeeld dat de rol van de bewerkbaarheid van de bodem in het verleden veel belangrijker was dan tegenwoordig. Het gevolg is dat een 'Romeinse' landevaluatie er anders uit komt te zien dan een moderne landevaluatie op basis van dezelfde informatie.

HOOFSTUK 10

Tot slot wordt in hoofdstuk 10 een aanzet gegeven tot het integreren van sociaal-culturele factoren in archeologische voorspellingsmodellen. Hoewel het landschap een zeer belangrijke factor is bij de voorspelling van de locatie van archeologische vindplaatsen, kan het landschap alleen niet een volledige verklaring bieden voor de vraag waarom de mens zich ergens wel of niet vestigde. Sociale en culturele factoren, zoals de nabijheid van andere nederzettingen, of de aanwezigheid van cultusplaatsen, hebben hierin ook een rol gespeeld. Dit feit is altijd aangevoerd als een belangrijk bezwaar tegen de huidige voorspellingsmodellen. Opvallend is echter dat er vrijwel geen voorbeelden bekend zijn waarin wordt geprobeerd om alsnog de sociaal-culturele component in de modellen te betrekken. In dit hoofdstuk wordt kort de problematiek toegelicht, en worden een aantal richtingen aangegeven waarmee sociaal-culturele factoren kunnen worden toegevoegd. Hierbij wordt niet uitgegaan van vage concepten zoals de prehistorische perceptie van het landschap, maar van meetbare aspecten, zoals de toegankelijkheid van een gebied tot bekende vindplaatsen, de zichtbaarheid van landschapselementen die mogelijk een belangrijke symbolische functie hadden, en de mate waarin er continuïteit van bewoning is geweest. Een dergelijke aanpak is praktisch gezien mogelijk, en leidt tot voorspellingsmodellen die niet alleen wetenschappelijk beter gefundeerd zijn, maar ook tot betere voorspellingen zullen leiden.

CONCLUSIES

Tenslotte wordt aan het eind van het proefschrift een poging gedaan om de toekomst van archeologische verwachtingskaarten in Nederland te schetsen. Er zijn drie scenario's denkbaar. In het eerste geval wordt er op de huidige voet doorgegaan, en blijven verwachtingskaarten een belangrijke rol spelen in de archeologische monumentenzorg, zonder dat bekend is wat de betrouwbaarheid van deze kaarten is. Het gevaar van deze ontwikkeling is dat er besluiten worden genomen over het al dan niet uitvoeren van inventariserend onderzoek, zonder dat duidelijk is welke risico's daaraan kleven. Deze risico's zijn zowel van archeologische aard (het 'missen' van interessante vindplaatsen omdat zij zich in een zone van lage verwachting bevinden), als van financiële aard. Het uitvoeren van inventariserend onderzoek dat weinig 'archeologisch rendement' oplevert is niet alleen duur, maar ondergraaft uiteindelijk ook het maatschappelijk draagvlak voor de archeologie.

Voor het verminderen van het archeologisch risico heeft het de voorkeur om helemaal geen verwachtingskaarten te gebruiken, maar gewoon altijd inventariserend onderzoek uit te voeren. Deze benadering, die in Frankrijk en Groot-Brittannië gebruikelijk is, zal in de praktijk niet uitvoerbaar zijn,

vanwege de financiële consequenties. In de praktijk zullen er altijd keuzes moeten worden gemaakt over waar wel of niet onderzoek wordt uitgevoerd. Deze keuzes kunnen beter gebaseerd zijn op een goed onderbouwde verwachtingskaart, dan op het subjectieve oordeel van individuele experts en/of beleidsmakers.

De derde optie, en tevens de eindconclusie van dit onderzoek, is dan ook dat voorspellingsmodellen moeten worden gemaakt die kwantitatieve technieken gebruiken om tot een schatting van de betrouwbaarheid van de voorspellingen te komen. Daarmee wordt het mogelijk om beter onderbouwde keuzes te maken over het wel of niet uitvoeren van onderzoek. Aan deze optie hangt echter wel een prijskaartje, want het vergt veel meer en complexere analyse van de archeologische gegevens dan tot op heden gebruikelijk is.