Predictive models of archaeological site distributions in New Zealand

SCIENCE & RESEARCH INTERNAL REPORT 181

J.R. Leathwick

Published by
Department of Conservation
P.O. Box 10-420
Wellington, New Zealand

Science & Research Internal Reports are written by DOC staff or contract science providers on matters which are on-going within the Department. They include reports on conferences, workshops, and study tours, and also work in progress. *Internal Reports* are not normally subject to peer review.
Publication was approved by the Manager, Science & Research Unit, Science Technology and Information Services, Department of Conservation, Wellington.
This report was prepared by J.R. Leathwick, Landcare Research, Hamilton, funded as part of Science & Research Investigation 3271.

ISSN 0114-2798
ISBN 0-478-21969-5

Cataloguing-in-Publication data
Leathwick, John R., 1954-
Predictive models of archaeological site distributions in New Zealand / J.R. Leathwick. Wellington, N.Z. : Dept. of Conservation, 2000.
1 v. ; 30 cm. (Science & Research internal report, 0114-2798 ; 181).
Includes bibliographical references.
ISBN 0478219695
1.Archaeological surveying--Mathematical models.
2. Archaeology--New Zealand. I. Title. Series: Science and Research internal report ; 181.

CONTENTS

Abstract

Predictive models provide a tool for distinguishing between locations where archaeological sites are likely to be present and locations where they are likely to be absent. This paper provides a preliminary analysis, using logistic regression analysis, of the environmental correlates of Maori pa and pit site locations in New Zealand. Predicted probabilities of occurrence are highest in high insolation, warm, summer-dry locations, mostly in northern and eastern New Zealand, with strong increases in probability in close proximity to major water bodies. Very low probabilities are predicted throughout the South Island, with the exception of sites close to the coast or rivers from Banks Peninsula northwards and west as far as about Farewell Spit. Technical issues relating to the adequacy of the data and the scale employed are also discussed.

1. Introduction

In late 1999, following a discussion between John Leathwick (Landcare Research), and Bruce McFadgen (Department of Conservation, DOC), an offer was made to carry out a preliminary analysis for DOC of the environmental correlates of Maori pa and pit site locations in New Zealand. The main objective was to explore the feasibility of predicting the likelihood of occurrence of known archaeological sites, with a view to providing guidance as to the degree of necessity for site inspections for proposed land developments.

The proposition that the locations of archaeological sites were strongly biased towards particular environments seemed reasonable, given the reliance of Maori in the pre-European era on cultivated foods such as kumara, and protein sources such as fish, both marine and freshwater. This suggested that correlations were likely both with climatic factors controlling plant growth (solar radiation, temperature, water supply and demand, and soil parent material), and with geographic parameters such as proximity to major water bodies.

2. Material and methods

2.1 DATA

A total of 11,251 locations of pa and pit sites were supplied by DOC from the Central Index of New Zealand Archaeological Sites (CINZAS) database that they administer. CINZAS provides an electronic index to the more than 53,000 records in the New Zealand Archaeological Association Site Recording Scheme.

Locations were specified as New Zealand Map Grid (NZMG) eastings and northings, but elevations were lacking. This was overcome by overlaying the geographic locations onto a 25-m resolution digital elevation model fitted by Landcare Research staff to Land Information New Zealand (LINZ)'s digital New Zealand Map Series (NZMS) 260 contour data.

Locations, now specified as easting, northing and elevation, were used as a basis for estimation of a range of climatic parameters, shown in previous work by Landcare Research staff to be relevant to the landscape scale distributions of biota (e.g. Leathwick 1998; Leathwick et al. 1998). These were:

- Mean annual temperature (MAT in °C).
- Temperature seasonality (TSEAS in °C), i.e. a measure of winter harshness calculated as the deviation in winter minimum temperatures for each site from that expected given its annual temperature.
- Mean annual solar radiation (MAS in MJ m^{-2} day^{-1}).
- Solar radiation seasonality (SSEAS in MJ m^{-2} day^{-1}), i.e. a measure comparing winter and annual solar radiation as for temperature seasonality.
- Ratio of rainfall to potential evapotranspiration in the driest month (MinR2PET, a ratio).
- Spring air saturation deficit (OctVPD in kPa), i.e. a strong determinant of evaporative demand.

Pa and pit site locations were also overlaid onto a digital copy of the New Zealand Land Resource Inventory (NZLRI) database to determine likely soil parent material. Fifteen soil parent material classes were identified from a version of the NZLRI 'toprock' data, modified to take better account of likely nutrient availability for plant growth. Finally, a GIS layer was created which indicates the geographic distance to major water bodies, i.e. the coast, and major lakes and rivers. Pa and pit site locations were overlaid onto this data layer to determine their proximity to major water bodies (WDIST).

2.2 ANALYSIS

The resulting matrix of site by environmental data indicating the presence of known pa and pit sites was imported into computer software package S-Plus. To enable the likelihood of occurrence of pa and pit locations to be assessed as a function of environment, these 'presence' data were combined with an equivalent-sized set of 'pseudo-absence' points, randomly selected from a 1-km grid across New Zealand. The logistic regression subsequently fitted to this combined dataset indicates the degree of bias of pa and pit site locations towards or away from particular environments, with values above 0.5 indicating a positive bias and lower values indicating a negative bias.

The logistic regression used for this analysis was fitted using Generalized Additive Models (Hastie & Tibshirani 1990). GAMs represent a recent refinement of a group of generalised regression techniques which offer advantages in their ability to handle non-normally distributed data, including the presence/absence data analysed here. GAMs offer advantages in that relationships between the response and continuous predictor variables are defined from the data using scatter-plot smoothers, rather than more inflexible parametric terms used traditionally for such analysis. The regression was fitted using a backwards stepwise procedure in which all variables were fitted in an initial model, and the significance of dropping each variable was tested in turn. Because of the highly skewed frequency distributions of MinR2PET and WDIST, these variables were subject to $\log_{10}$ transformation before analysis (Fig. 1). The contribution of individual terms in the final model was assessed using the change in residual deviance when dropping each term, with these changes distributed approximately as for a Chi-square statistic. A map showing predicted likelihood of occurrence was then produced using an equivalent set of environmental data for points on a 1-km grid across New Zealand.

3. Results

The final regression model contained all the terms initially fitted (Table 1), i.e. all make a statistically significant contribution to the outcome. The fitted terms also explain a substantial amount of the total deviance, although it should be noted that there is no single summary statistic (e.g. analogous to an R^2) available for use with the generalised regression technique used here.

Using the change in residual deviance when dropping each term from the final model as a measure, soil parent material and distance to major water bodies make the biggest contributions to the final model, closely followed by mean annual temperature (Table 2). Other factors are generally much less important, but are still highly statistically significant, reflecting in part the very large sample size used in the analysis. Examination of the fitted regression functions for the individual predictors indicates that pa and pit sites occur most frequently in environments in close proximity to water bodies (within 1 km), and having warm temperatures (14–15°C), high solar radiation (> 14.5 MJ m^{-2} day^{-1}), mild winters, and dry summers (OctVPD > 0.5 and MinR2PET < 1). Limestone and andesite are the parent materials having the highest levels of occurrence, and peat, alluvium, and ultramafic parent materials have the lowest levels of occurrence (Fig. 2).

The predictive map (Fig. 3) produced from the final regression clearly shows the influence of the environmental predictors described above, i.e. predicted probabilities of occurrence are highest in high insolation, warm, summer-dry locations, mostly in northern and eastern New Zealand, with strong increases in

TABLE 1. SUMMARY OF THE FINAL REGRESSION MODEL.

	RESIDUAL DEVIANCE	DEGREES OF FREEDOM
Null model	30.6×10^3	22,096
Final regression model	9.6×10^3	22,054

probability in close proximity to major water bodies. Very low probabilities are predicted throughout the majority of the South Island, with the exception of sites close to the coast or rivers from Banks Peninsula northwards and west as far as about Farewell Spit.

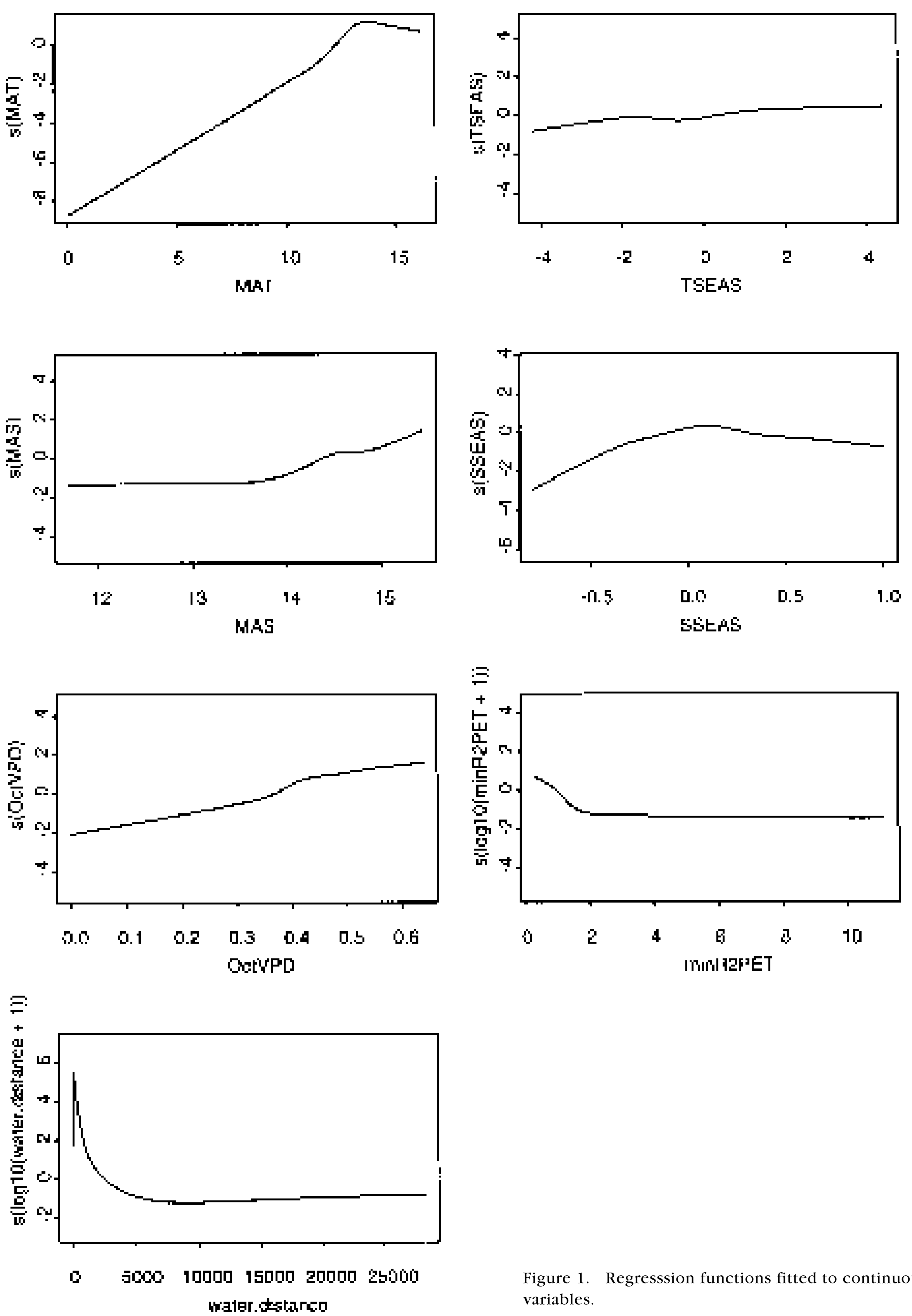

Figure 1. Regresssion functions fitted to continuous predictor variables.

TABLE 2. INCREASES IN RESIDUAL DEVIANCE, AND THEIR STATISTICAL SIGNIFICANCE, WHEN DROPPING TERMS FROM A REGRESSION RELATING PA AND PIT SITE LOCATIONS TO ENVIRONMENT.

VARIABLE	CHANGE IN DEVIANCE	DEGREES OF FREEDOM	SIGNIFICANCE
Mean annual temperature	675.2	3.6	0
Temperature seasonality	58.5	3.8	$4.62e^{-12}$
Mean annual solar radiation	232.2	3.9	0
Solar radiation seasonality	215.4	3.6	0
October VPD	274.6	3.8	0
Minimum rainfall/PET	65.9	3.7	$1.03e^{-13}$
Distance to major water	744.9	4.4	0
Soil parent material	767.6	14	0

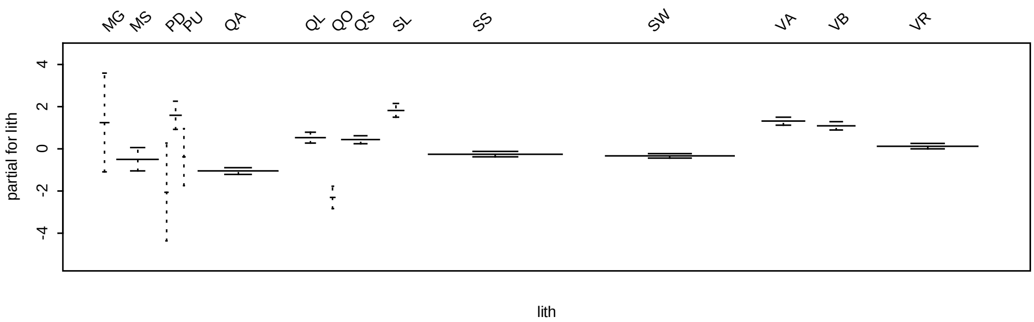

Figure 2. Regression coefficients fitted to soil parent materials. Vertical dashed lines show standard errors, whereas the width of the central bar is proportional to the number of cases in each parent material category. Codes in the same order as on the graph are: MG—gneiss, MS—schist, PD—diorite, PG—granite, PU—ultramafic, QA—alluvium, QL—loess, QO—organic, QS—sand, SL—limestone, SS—strong sedimentary, SW—weak sedimentary, VA—andesite, VB—basalt, VR—rhyolite.

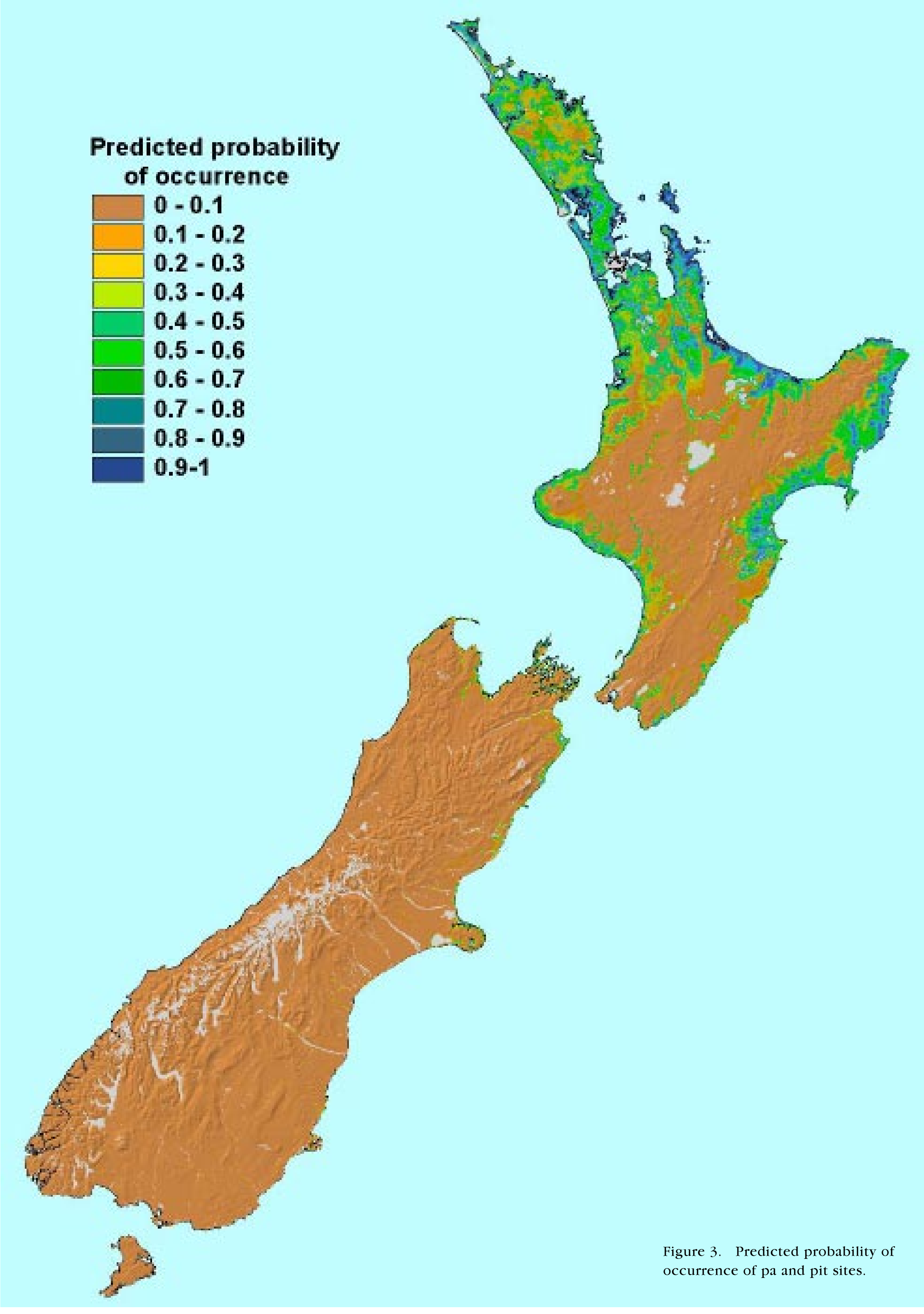

Figure 3. Predicted probability of occurrence of pa and pit sites.

4. Discussion: Adequacy of the analysis

4.1 TECHNICAL CHALLENGES

The analysis presented two technical challenges that required the development of appropriate work-arounds, as follows:

1. The absence of elevations in the original CINZAS database records meant that these had to be estimated by overlaying the plot locations onto a digital elevation model. Given that the locations are at best accurate to only 100 m along both eastings and northings, the elevation estimates may not be reliable for fine-scale work, particularly in steep coastal locations often preferred for pa sites. Although this is unlikely to substantially affect the subsequent climate estimates, caution should be exercised in interpretation of the raw elevations (or slopes), e.g. if they were to be used as a measure of the topographic preferences of pa and pit site locations.
2. The presence-only nature of the pa and pit site data as supplied placed some inherent limitations on the range of possible analysis approaches. Normally such data can only be analysed to show either the environmental range of occurrences, or the density (counts) of occurrences in particular environments. Such approaches are much less informative than analyses that relate the relative incidence of presences and absences to environment, with the analysis then able to indicate the probability of occurrence in particular environments. In this analysis an approximation of a presence/absence analysis was made by including an equivalent-sized, random selection of points from New Zealand's environment, which were treated as absences. This has two consequences. First, some of these pseudo-absence points may in fact coincide with the presence of pa and pit sites, and for these points the probability of occurrence will be underestimated for the environments they represent. Second, the probability values fitted by the regression are only relative, and lack any absolute meaning. They should however, behave in an ordinal fashion, so that higher fitted values can be confidently regarded as indicating a higher relative likelihood of occurrence.

4.2 ISSUES OF SCALE

The trial regression analysis carried out for this pilot study has been used to make predictions only at very broad spatial scales, i.e. at a grid resolution of 1 km. Given the inherent accuracy of the underlying data, this could probably be increased safely up to a grid resolution of 20 m, i.e. to a level suitable for mapping at spatial scales of 1 : 50,000. However, this would require careful consideration of the role of distance to water in the analysis and subsequent prediction. Use could be made of digital stream network data available from LINZ, either from 1 : 50,000 or 1 : 250,000 scale mapping. This would probably still require 'major' water bodies to be identified, as otherwise the myriad of minor streams would confuse the picture shown by this analysis.

An alternative, and probably more cost-effective, approach would be to transfer results from this (or preferably a more thorough) analysis into an Excel spreadsheet or ArcView application. Probabilities of occurrence could then be determined by first overlaying co-ordinates for a site of interest onto paper or digital copies of climate maps to determine site climate parameters. These could then be either manually or automatically passed to a macro which would calculate the likelihood of pa or pit site occurrence for that particular site. Obviously such analyses could also be refined to differentiate between different types of archaeological sites.

5. Conclusions

- Results from this pilot analysis indicate a considerable potential to identify sites likely to have archaeological significance, based on their environmental attributes.
- Although existing archaeological site data are essentially 'presence-only' in character, use of pseudo-absence data allows the relative likelihood of pa and pit site occurrence to be estimated.
- Both climate and landform factors are important in determining the likelihood of occurrence of pa and pit sites, with greatest likelihood of occurrence coinciding with warm, sunny, summer-dry environments, particularly those combining limestone, granite, andesite, or basalt parent materials with close proximity to major water bodies.
- Implementation of this analysis to enable guidance of site inspections for proposed land developments could be achieved by mapping predicted probability or occurrence at scales up to a maximum resolution of 1 : 50,000.
- Alternatively a computer-based expert system could be developed allowing prediction of the likelihood of occurrence on a site-by-site basis.

6. References

Hastie, T.; Tibshirani, R. J. 1990: *Generalized Additive Models*. Chapman and Hall, London.

Leathwick, J. R. 1998: Are New Zealand's *Nothofagus* species in equilibrium with their environment? *Journal of Vegetation Science 9:* 719–732.

Leathwick, J. R.; Burns, B. R.; Clarkson, B. D. 1998: Environmental correlates of tree alpha-diversity in New Zealand primary forests. *Ecography 21:* 235–246.