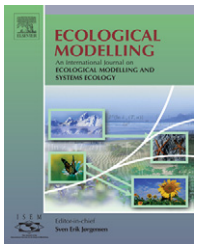
available at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/ecolmodel



Analysis of environmental factors influencing the range of anopheline mosquitoes in northern Australia using a genetic algorithm and data mining methods

A.W. Sweeney[a,*], N.W. Beebe[a,**], R.D. Cooper[b]

[a] Institute for the Biotechnology of Infectious Diseases, University of Technology Sydney, PO Box 123, Broadway, NSW 2007, Australia
[b] Army Malaria Institute, Gallipoli Barracks, Enoggera, Qld 4052, Australia

ARTICLE INFO

Article history:
Received 20 September 2005
Received in revised form
2 December 2006
Accepted 11 December 2006
Published on line 16 January 2007

Keywords:
Ecological niche modelling
Genetic algorithms
Data mining
Mosquitoes
GARP
CART
KnowledgeSeeker

ABSTRACT

Environmental factors which influence the distributions of malaria vectors in northern Australia (*Anopheles farauti* ss, *A. farauti* 2 and *A. farauti* 3) were investigated by ecological niche modelling and data mining using an extensive data set of species presence and absence records obtained by systematic field surveys. Models were generated with GARP (the genetic algorithm for rule-set prediction) using geographical coverages of 41 climatic and topographic parameters for the north of the continent. Environmental variables associated with species records were identified with the ranking procedures of the decision tree software packages CART and KnowledgeSeeker. There was consistent agreement in the variables ranked by both methods. This permitted the selection of reduced sets of environmental variables to develop GARP models for the three target species with equivalent predictive accuracy to those which used all of the environmental information. The environmental parameters which define the realised distributions of *A. farauti* ss and *A. farauti* 3 were well described by this approach but the results were less satisfactory for *A. farauti* 2. Atmospheric moisture was shown to be a critical variable for each species which accords with many field and laboratory observations concerning the influence of humidity on adult mosquito survival.

1. Introduction

Members of the *Anopheles punctulatus* group are important malaria vectors in the southwest Pacific region. Understanding their geographical ranges is of considerable interest from both ecological and disease transmission viewpoints. Comprehensive surveys of the Australian species of the group, *Anopheles farauti* ss, *A. farauti* 2 and *A. farauti* 3, were undertaken in the formerly malarious areas of northern Australia between 1985 and 1994. The surveys, which included material from over 600 localities, yielded more than 300 locality records of the target species to provide a realistic indication of their realised distribution in the Northern Territory and Queensland. The results showed that these mosquitoes were distributed around the coast and 50–100 km inland north of 20°S latitude and east of 129°E longitude. In some areas the three species were found together but the overall patterns of occurrences for each species were different (Sweeney et al., 1990; Cooper et al., 1995, 1996). These apparent dissimilarities in realised distribution suggest that there may be differences in the ecological factors which influence the range of the individual species.

* *Corresponding author*. Tel.: +61 2 4385 8774; fax: +61 2 9514 4003.
** *Corresponding author*.
E-mail address: tony.sweeney@uts.edu.au (A.W. Sweeney).
0304-3800/$ – see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.ecolmodel.2006.12.003

Ecological niche modelling with GARP (the genetic algorithm for rule-set prediction) has proved to be particular useful for investigating species distributions by creating models, using point localities where species are known to occur and environmental data for the geographic region of interest, to predict species range (Stockwell and Peters, 1999). A comparative study of GARP predicted ranges for the five sibling species of *Anopheles quadrimaculatus* implied that *A. quadrimaculatus* ss was the only member of the group capable of transmitting malaria throughout the formerly malarious area of the United States (Levine et al., 2004a). Similarly, Levine et al. (2004b) used GARP to predict the potential range of the *Anopheles gambiae* complex in Africa. Models developed from African distribution data and projected to South America suggested that *A. gambiae* ss was the species introduced into Brazil in 1929.

For the present study we used GARP for modelling the potential distribution of the *A. farauti* sibling species in Australia. Inputs for model construction included species occurrence data from field surveys together with high resolution environmental information for northern Australia based on historical climate records. There are a number of procedures other than GARP which have been used for predictive modelling of species distributions including regression algorithms (such as generalised linear models and locally weighted regression), classification or decision tree analysis, environmental envelopes (including BIOCLIM which is part of the GARP algorithm), neural networks, and Bayesian statistics (reviewed by Guisan and Zimmerman (2000)). In addition to ecological modelling some of these analytical procedures can be applied to data mining, the process of statistical analysis to reveal previously unknown patterns from a set of data values. We selected two decision tree software packages, CART (classification and regression tree analysis, Breiman et al., 1984) and KnowledgeSeeker (Biggs et al., 1991), to search for significant environmental factors associated with species presence. GARP models were generated with reduced sets of environmental layers highlighted by these data mining techniques to determine whether the model outputs corresponded with high quality range predictions. The overall objective was to identify the key environmental factors responsible for defining the geographical ranges of the different vector species as such factors are of major epidemiological significance and of direct relevance for malaria control strategies.

2. Materials and methods

2.1. *Mosquito surveys*

The area surveyed included the coast of Queensland and the Northern Territory and up to 300 km inland between latitudes 10–19°S and longitudes 128–146°E (Fig. 1a). A different sector of this region was covered progressively each year between 1985 and 1991: the Queensland coast from Townsville to Cooktown in 1985, Cape York Peninsula in 1986, the southern coast of the Gulf of Carpentaria in 1987 and the Northern Territory in 1988, 1989, and 1990. The 1991 survey included the Torres Strait Islands and areas of far north Queensland not adequately covered in previous surveys. In 1994 a final survey was made in the Northern Territory to provide additional

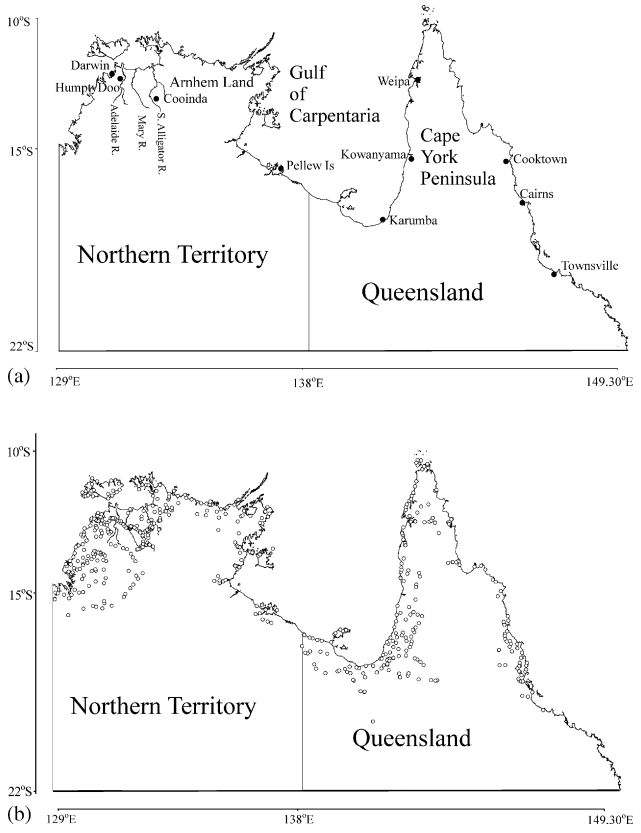


Fig. 1 – Survey area of northern Australia. (a) Place names and localities mentioned in text. (b) Localities of 619 collection sites (designated ○) at which anopheline mosquitoes were collected during field surveys. These included both record and no-record sites for *A. farauti* sl.

material in the floodplains of the Adelaide and Mary Rivers. Each yearly survey was made for a month during March–May which normally coincides with the end of the north Australian wet season when larval sites are expected to be plentiful and adult densities are usually high. In order to adequately sample the major land cover classes in this thinly populated part of the world the surveys relied on Australian Army helicopters to access remote areas inaccessible by roads and tracks. Adult mosquitoes were caught overnight in battery-powered CO_2 baited light traps (Rohe and Fall, 1979). Each locality was plotted on 1:100,000 Australian Army Topographical maps which show grid lines at 1000 m intervals. The latitude and longitude of individual collection sites were plotted to a tenth of a grid square resulting in a spatial accuracy to the nearest 100 m. Specimens were morphologically identified to species level in the field and the collections from each locality were then stored separately in liquid nitrogen for later identification of sibling species in the laboratory using either allozyme electrophoresis (Mahon, 1984) or DNA probes (Cooper et al., 1991). Anopheline larvae, collected from a wide range of water bodies, were reared in a field laboratory to the adult stage and then processed in the same way as adult specimens. The survey data set comprises collection records at 619 localities at which members of the *A. farauti* group were either found or not found (Fig. 1b). There were over 26,000 anopheline mosquitoes

collected in these surveys of which 5482 were identified as *A. farauti* sl.

2.2. Environmental layers

Raster ASCII grids were generated for northern Australia at a spatial resolution of 0.01° (approximately 1 km) with geographical extents of 10–22°S latitude and 128–150°E longitude for 41 environmental layers (see Table 1). These included 27 climatic parameters for temperature, rainfall and radiation (p1–p27) produced by BIOCLIM using the ANUCLIM software package (Houlder et al., 1999). This procedure involved the use of monthly mean climate surface coefficients, generated by the thin plate smoothing spline technique ANUSPLIN (Hutchinson, 2003), from Australian Bureau of Meteorology climate data between 1921 and 1995 (Hutchinson and Kesteven, 1998). The geographic coordinates of the meteorological stations were used as independent spline variables together with a 0.01° digital elevation model (DEM) for northern Australia generated with ANUDEM (Hutchinson, 1997) which acted as a third independent variable. Atmospheric moisture is known to be an important factor influencing survival and longevity of adult mosquitoes so we included four climatic layers for dewpoint (for January and July at 9 a.m. and 3 p.m.) generated with ESOCLIM (a component of ANUCLIM) and six layers for relative humidity (for January, July and annual mean at 9 a.m. and 3 p.m.) generated with ANUSPLIN from long term climate data obtained from the Australian Bureau of Meteorology. We used four non-climatic parameters: elevation (the 0.01° DEM generated with ANUDEM); slope and aspect (both generated from the elevation DEM) as well as distance from the coast—an ASCII grid with cell values representing kilometres from the coastal outline of northern Australia generated with the GIS program TNTmips (MicroImages Inc., Lincoln, Nebraska).

2.3. Range modelling with GARP

DesktopGarp version 1_1_6 (Scachetti-Pereira, 2003)was used for this study. Models were generated using the record sites for each species as inputs together with various combinations of the environmental layers. The GARP procedure was implemented using half of the species record sites as a training data set for model building and the other half for model testing. The number of models generated for each GARP run and the optimisation parameters for each model can be preset to control when the algorithm stops. We generated 100 models for each species and specified optimisation parameters to limit the number of iterations per model to 1000 or when the convergence limit (i.e. the change between successive iterations) reached 0.01. The best subsets procedure (Anderson et al., 2003) was used to select 5 out of each set of 100 models. These were added together using TNTmips to produce predicted range maps for each species.

2.4. Data mining

The mosquito survey data in northern Australia include the geographic coordinates of over 600 localities at which particular species were either found (record sites) or not found (no-record sites). The values for the 41 environmental layers

Table 1 – Description of surfaces of northern Australia generated for environmental parameters

Parameter	Description
anntemp (p1[a])	Annual mean temperature. The mean of all weekly mean temperatures
diurntemp (p2[a])	Mean diurnal range. The mean of all weekly diurnal temperature ranges
isotemp (p3[a])	Isothermality. diurntemp (p2) divided by rangetemp (p7)
seastemp (p4[a])	Temperature seasonality (coefficient of variation[b])
maxtemp (p5[a])	The highest temperature of the weekly maximum temperatures
mintemp (p6[a])	The lowest temperature of the weekly minimum temperatures
rangetemp (p7[a])	Temperature annual range: maxtemp–mintemp
tempwetqtr (p8[a])	Mean temperature of the wettest quarter
tempdryqtr (p9[a])	Mean temperature of the driest quarter
tempwarmqtr (10[a])	Mean temperature of the warmest quarter
tempcoldqtr (p11[a])	Mean temperature of the coldest quarter
annrain (p12[a])	Annual precipitation
rainwetmth (p13[a])	Precipitation of wettest month
raindrymth (p14[a])	Precipitation of driest month
seasrain (p15[a])	Precipitation seasonality (coefficient of variation[b])
rainwetqtr (p16[a])	Precipitation of wettest quarter
raindryqtr (p17[a])	Precipitation of driest quarter
rainwarmqtr (p18[a])	Precipitation of warmest quarter
raincoldqtr (p19[a])	Precipitation of coldest quarter
annrad (p20[a])	Annual mean radiation. The mean of all weekly radiation estimates
highrad (p21[a])	The largest weekly radiation estimate
lowrad (p22[a])	The lowest weekly radiation estimate
seasrad (p23[a])	Radiation seasonality (coefficient of variation[b])
radwetqtr (p24[a])	Radiation of wettest quarter
rad_dryqtr (p25[a])	Radiation of driest quarter
radwarmqtr (p26[a])	Radiation of warmest quarter
radcoldqtr (p27[a])	Radiation of coldest quarter
dp9jan	Dewpoint for January at 9 a.m.
dp9jul	Dewpoint for July at 9 a.m.
dp3jan	Dewpoint for January at 3 p.m.
dp3jul	Dewpoint for July at 3 p.m.
rh9ann	Mean annual relative humidity at 9 a.m.
rh9jan	Relative humidity for January at 9 a.m.
rh9jul	Relative humidity for July at 9 a.m.
rh3ann	Mean annual relative humidity at 3 p.m.
rh3jan	Relative humidity for January at 3 p.m.
rh3jul	Relative humidity for July at 3 p.m.
coastdist	Distance from the coast (km)
elev	Elevation above sea level (m)
slope	Slope of land surface (degrees above horizontal)
aspect	Aspect of land surface (compass degrees)

[a] BIOCLIM parameter number.
[b] The coefficient of variation (for temperature, precipitation or radiation) is the standard deviation of the weekly mean estimates expressed as a percentage of the mean of those estimates in Kelvin.

at grid cells corresponding to points representing record sites and no-record sites for *A. farauti* ss, *A. farauti* 2 and *A. farauti* 3 in northern Australia were generated with TNTmips. This information was then transferred to database tables which contained a record/no-record field for each species as well as separate fields containing the values for each of the environmental layers. The record/no-record field was assigned as the dependent variable for decision tree analysis to explore relationships among the environmental layers which constitute the independent variables.

Both CART and KnowledgeSeeker split records in the dependent variable and display statistically significant patterns among the independent variables. CART makes binary splits of the independent variables to construct a decision tree. One output of the CART procedure considers the importance of the independent variables which are ranked in descending order of their contribution to tree construction. This is not determined solely by primary splits as CART keeps track of surrogate splits in the tree-growing process (Steinberg and Colla, 1995). To calculate a variable importance score CART looks at the improvement measure attributable to each variable in its role as a surrogate to the primary split. The values of these improvements are summed over each node of the tree and scaled relative to the best performing variable. The variable with the highest sum of improvements is scored 100, and all other variables have lower scores ranging downwards towards zero. KnowledgeSeeker examines the data in each of the independent variables and searches for relationships with the dependent variable. Variables of most importance can be presented in order of statistical significance using chi-square analysis. The Important Splits command ranks the variables in descending order from the most significant split. We used the variable importance ranking procedures of CART and KnowledgeSeeker to indicate the main environmental variables identified by each as being associated with presence or absence of the target mosquito species. The geographical coordinates of the collection localities were not included among the independent variables so the suite of climatic and topographic variables were ranked empirically without considering their spatial context.

2.5. *Procedures for identifying key range variables*

The numbers of mosquitoes collected from different survey sites ranged from 1 to 3500. No-record sites which yielded <10 anopheline specimens were excluded from analysis with CART and KnowledgeSeeker in order to reduce sampling bias. This decision was based on our long term observations in the malaria receptive area of northern Australia which indicate that other north Australian *Anopheles* species are usually found in collections of *A. farauti* sl. We considered that a collection of >9 anophelines which did not include a member of the *A. farauti* group was a reasonable lower limit for inclusion as a no-record site in data mining procedures. Thus the dependent variable for our initial data mining iterations included all of the record sites for each species together with the no-record sites from collections of >9 specimens (data set 1). The first series of GARP runs included all 41 environmental layers to generate baseline range maps for each species. When data set 1 was overlaid on the baseline GARP models it was apparent that a significant number of no-record sites were within the predicted range for each species. These sites were excluded from a second round of data mining iterations. For this series, designated data set 2, the dependent variable included all of the record sites for each species (as for data set 1) but only the no-record sites >9 specimens outside the baseline predicted range (based on all of the environmental layers) for that species.

GARP models were made with the most significant environmental layers identified for each species by data mining in data set 2. These included all of the variables ranked for each species by either CART or KnowledgeSeeker. A final series of models were produced for each species to investigate whether the number of environmental layers could be further reduced without impairing the quality of range predictions. Layers were identified for removal with the N-1 jackknife procedure (Peterson and Cohoon, 1999). This involved the sequential removal of separate single layers from the environmental data set for a series of GARP runs. The results were scored by ranking those layers whose removal had the greatest negative effect on test accuracy. As analysis of all combinations of the 41 environmental layers by this procedure would have been prohibitive, the N-1 jackknife method was employed for the reduced set of variables identified by data mining. One different ranked variable was removed sequentially from successive GARP runs and the range predictions were compared with the baseline range prediction which included all of the ranked variables for that species. For example, with 15 variables ranked by either CART or KnowledgeSeeker this involved separate GARP runs with 14 environmental layers in which a different layer was excluded each time. The range maps derived from the best subsets of 100 models for each run were then compared with that of the GARP run using all layers to identify the ones whose removal resulted in the largest discrepancies from the baseline predicted range using all of the environmental information. The range predictions were also assessed in relation to the known distribution of the species obtained from the survey results.

3. Results

3.1. *Range models based on all environmental layers*

The best subset models derived from GARP runs using 41 environmental layers were in good agreement with the observed distributions for each species as all of the survey record sites (127 for *A. farauti* ss, 68 for *A. farauti* 2, and 120 for *A. farauti* 3) were within the baseline predicted range maps for the individual species. For *A. farauti* ss, the only member of the group with an exclusively coastal distribution, the range map followed the coastal trend in the Northern Territory as well as around the perimeter of the Cape York Peninsula and along the eastern coast of north Queensland without extending inland to areas where it has not been found (Fig. 2a). It incorporated the central portion of the southern coast of the Gulf of Carpentaria adjacent to a record site in the Pellew Islands but not further east along the Gulf coast near Karumba where members of the *A. farauti* group were not collected. However, the predicted range included the entire western coast of Cape York Peninsula further south than the record sites of this species. The

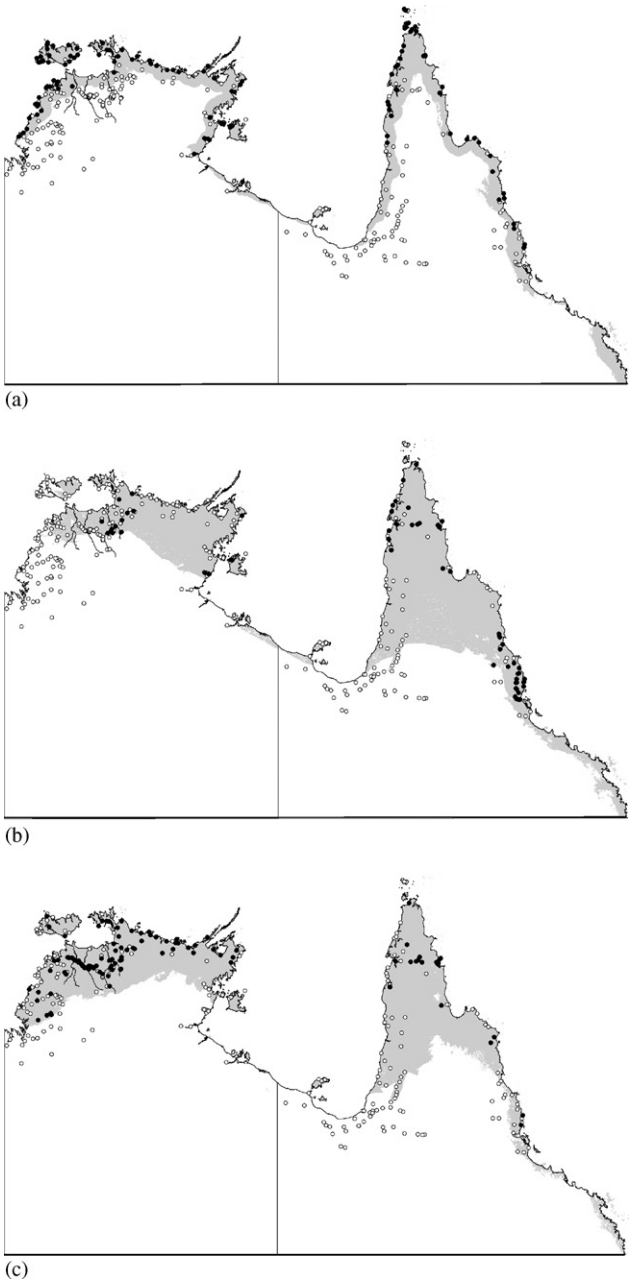


Fig. 2 – Baseline predicted range maps based on best subsets of 100 GARP models with all 41 environmental layers. (●) record sites; (○) no-record sites with ≥10 specimens. (a) *A. farauti* ss. (b) *A. farauti* 2. (c) *A. farauti* 3.

equivalent Garp models for *A. farauti* 2 resulted in a range prediction which was only fair as it included the southern coast of the Gulf of Carpentaria and the entire west coast of Cape York Peninsula in Queensland as well as an area south and west of Darwin in the Northern Territory where this mosquito has not been found (Fig. 2b). The best subset models for *A. farauti* 3 encompassed the record localities for this species in the Northern Territory and Queensland without including areas where this species is thought to be absent (Fig. 2c). Unlike the range predictions for *A. farauti* 2 it correctly excluded the southern coast of the Gulf of Carpentaria but in common with the other two species it extended along the west coast of Cape York Peninsula beyond the range of the record sites.

3.2. *Data mining analysis of mosquito data sets*

The first data mining iterations were made with data set 1 which included record sites for each species together with the no-record sites from collections of 10 or more specimens. The same variables were identified by both CART and KnowledgeSeeker in 7/10 significant splits of *A. farauti* ss, in 4/10 splits of *A. farauti* 2 and in 5/10 splits of *A. farauti* 3 (Table 2). When data set 1 was overlaid on the baseline GARP models (Fig. 2a–c) it was observed that many of the no-record sites were inside the predicted range for each species. These varied from 53 of 231 no-record sites for *A. farauti* ss, 108 of 269 no-record sites for *A. farauti* 2 and 111 of 221 no-record sites for *A. farauti* 3. These sites were excluded from data set 2 which included all of the record sites for each species (as for data set 1) but only the no-record sites of >9 specimens outside the baseline GARP range predictions (using the full complement of 41 environmental variables). The data mining results with data set 2 showed that there were 7/10 common variables ranked by both CART and KnowledgeSeeker for *A. farauti* ss and 5/10 common variables for *A. farauti* 2 (Table 2). CART identified only 8 significant environmental variables among data set 2 for *A. farauti* 3. Seven of these were among the 10 variables ranked by KnowledgeSeeker for this species.

Inspection of the statistical outputs for CART showed that, for all three mosquito species, there was a marked increase in classification success rates of data set 2 for presence and absence records compared with those of data set 1 (Table 3). This improvement was most significant for *A. farauti* 3 in which the predicted presence/actually present success rate increased from 78.3% in data set 1 to 97.3% in data set 2. The χ^2 values for variables ranked by KnowledgeSeeker for *A. farauti* ss and *A. farauti* 3 were considerably larger in data set 2, indicating a higher level of statistical significance, than the equivalent values for the same variables in data set 1 (Table 2). On the other hand, the ten variables ranked for *A. farauti* 2 by KnowledgeSeeker had a similar numerical range of χ^2 values for both data sets.

3.3. *Model development with GARP*

Further range modelling with GARP was undertaken for each species using the most significant environmental layers identified in data set 2 by either CART or KnowledgeSeeker to produce range maps based on 13 layers for *A. farauti* ss (Fig. 3a); 15 layers for *A. farauti* 2 (Fig. 3b); and 11 layers for *A. farauti* 3 (Fig. 3c). In each case the predicted range was very similar to that from the baseline GARP series which used all of the environmental layers.

A final series of GARP models were made to explore the minimum number of environmental layers associated with good quality range predictions for each species. Different combinations of the layers ranked by the data mining methods were systematically selected for GARP runs using the jackknife procedure devised by Peterson and Cohoon (1999). Range maps were based on the best subsets of 100 GARP models for each species in which one different layer was removed each

Table 2 – Environmental variables ranked by CART and KnowledgeSeeker as being associated with mosquito species in data set 1[a] and data set 2[b] with dependent variable splits between record and no-record sites

Rank	Data set 1				Data set 2			
	CART		KnowledgeSeeker		CART		KnowledgeSeeker	
	Variable	Score	Variable	χ^2	Variable	Score	Variable	χ^2
Anopheles farauti ss								
1	coastdist[c]	100	coastdist[c]	201.6	coastdist[c]	100	coastdist[c]	263.2
2	rh3ann[c]	81.0	rh3jan[c]	150.3	rh3ann[c]	72.3	rangetemp[c]	190.3
3	rh3jul	69.2	rangetemp[c]	149.8	rangetemp[c]	65.3	rh3jan[c]	189.5
4	rh3jan[c]	67.7	rh3ann[c]	142.1	rh3jul[c]	65.3	rh3ann[c]	185.3
5	maxtemp[c]	61.0	dp3jul	141.3	maxtemp[c]	62.9	diurntemp[c]	174.2
6	elev	37.9	mintemp	133.7	rh3jan[c]	62.3	dp3jul	169.2
7	rangetemp[c]	9.6	dp9jul	130.5	diurntemp[c]	7.9	rh3jul[c]	168.5
8	diurntemp[c]	8.5	seastemp[c]	125.4	lowrad	2.3	maxtemp[c]	167.6
9	seastemp[c]	8.0	maxtemp[c]	124.9	rainwetqtr	2.3	dp9jul	162.2
10	seasrain	6.4	diurntemp[c]	124.5	raincoldqtr	2.3	mintemp	153.3
A. farauti 2								
1	lowrad[c]	100	radcoldqtr[c]	151.1	dp9jul[c]	100	lowrad[c]	159.3
2	annrad[c]	98.5	lowrad[c]	134.5	rh9jul[c]	82.4	radcoldqtr[c]	150.7
3	radcoldqtr[c]	93.8	annrad[c]	134.5	dp3jul[c]	77.2	annrad	140.5
4	radwetqtr[c]	89.9	radwetqtr[c]	134.5	rh3ann	70.5	radwetqtr	133.1
5	tempwarmqtr	88.7	rainwarmqtr	131.7	diurntemp	68.7	rad_dryqtr	131.4
6	tempwetqtr	69.2	seasrain	121.7	rangetemp	67.4	rh9jul[c]	124.5
7	anntemp	24.2	radwarmqtr	120.1	radcoldqtr[c]	35.7	dp9jul[c]	123.4
8	raincoldqtr	14.6	rainwetmth	117.7	lowrad[c]	35.7	rainwarmqtr	122.8
9	maxtemp	8.5	annrain	110.6	seasrad	33.9	dp3jul[c]	112.2
10	aspect	3.0	rainwetqtr	109.4	dp9jan	30.3	rainwetqtr	111.1
A. farauti 3								
1	seastemp[c]	100	seastemp[c]	101.3	seastemp[c]	100	annnrain[c]	189.8
2	annrain[c]	78.4	janrh9	100.3	annrain[c]	88.7	seastemp[c]	186.7
3	isotemp[c]	70.5	isotemp[c]	81.4	isotemp[c]	84.8	isotemp[c]	167.0
4	mintemp	62.0	rainwetqtr	73.7	tempcoldqtr[c]	70.6	rainwetqtr	155.3
5	tempcoldqtr[c]	59.0	tempcoldqtr[c]	68.8	dp3jan[c]	56.5	rh9jan[c]	148.4
6	dp3jan[c]	52.7	dp3jul	68.6	tempdryqtr[c]	55.4	tempcoldqtr[c]	147.9
7	coastdist	28.3	dp3jan[c]	66.8	rh9jan[c]	5.5	dp3jan[c]	126.2
8	elev	18.8	annrain[c]	66.7	radcoldqtr	1.75	rainwetmonth	124.8
9	dp9jan	11.0	rainwetmth	65.5			mintemp	115.0
10	rh3jan	3.8	rh9jul	64.8			tempdryqtr[c]	113.7

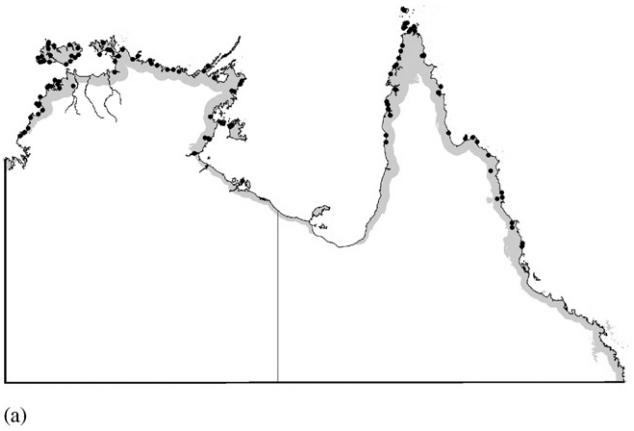
[a] Data set 1 contains all record sites of the species as well as no-record sites (with ≥10 specimens) within and outside the predicted range of the species based on the best subsets of 100 GARP models with all 41 environmental layers.
[b] Data set 2 contains all record sites of the species as well as no-record sites (with ≥10 specimens) outside the predicted range of the species based on the best subsets of 100 GARP models with all 41 environmental layers.
[c] Variables ranked by both CART and KnowledgeSeeker for the same species data set.

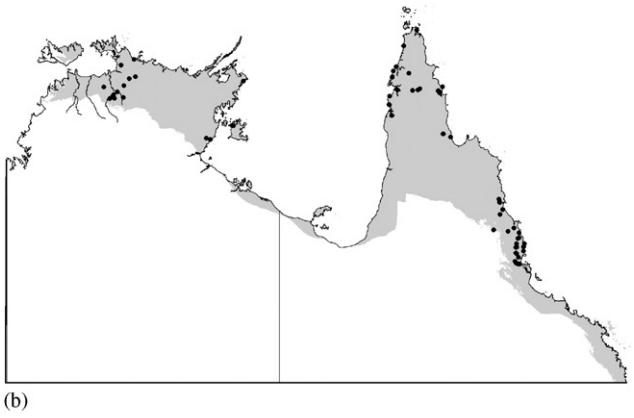
time from the ranked variables identified by either CART or KnowledgeSeeker in data set 2 (13 for *A. farauti* ss, 15 for *A. farauti* 2 and 11 *A. farauti* 3). The output maps for the different models were inspected to assess which removed layers had the most adverse effect on predicted range. Following this procedure a reasonably good range prediction for *A. farauti* ss was obtained from nine environmental layers which included the seven variables ranked by both CART and Knowledge-

Table 3 – Classification success rates for data sets using CART

Species	Data set 1		Data set 2	
	Present[a] (%)	Absent[b] (%)	Present[a] (%)	Absent[b] (%)
A. farauti ss	85.9	89.1	88.9	97.6
A. farauti 2	81.8	86.8	96.3	94.1
A. farauti 3	78.3	80.0	97.3	96.7

[a] Predicted present and actually present.
[b] Predicted absent and actually absent.





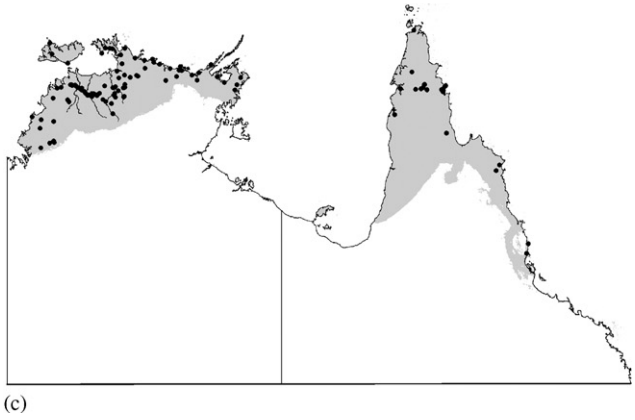
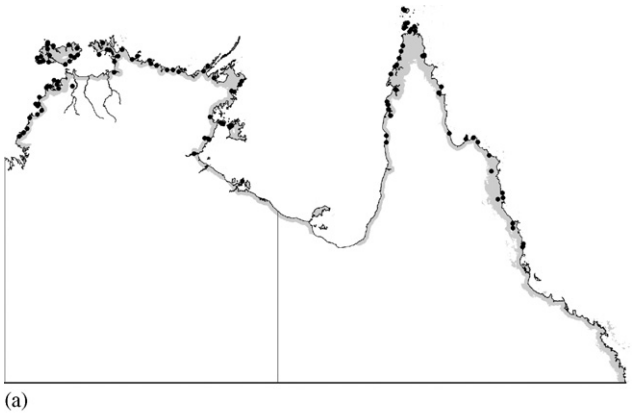


Fig. 3 – Predicted range maps based on significant variables ranked by either CART or KnowledgeSeeker in data set 2: (a) *A. farauti* ss based on 13 variables: coastdist, diurntemp (p2), maxtemp (p5), mintemp (p6), rangetemp (p7), rainwetqtr (p16), raincoldqtr (p19), lowrad (p22), dp3jul, dp9jul, rh3ann, rh3jan and rh3jul. (b) *A. farauti* 2 based on 15 variables for *A. farauti* 2: diurntemp (p2), rangetemp (p7), rainwetqtr (p16), rainwarmqtr (p18), annrad (p20), lowrad (p22), seasrad (p23), radwetqtr (p24), rad_dryqtr (p25), radcoldqtr (p27), dp3jul, dp9jan, dp9jul, rh9jul and rh3ann. (c) *A. farauti* 3 based on 11 variables: isotemp (p3), seastemp (p4), mintemp (p6), tempdryqtr (p9), tempcoldqtr (p11), annrain (p12), rainwetmonth (p13), rainwetqtr (p16), radcoldqtr (p27), dp3jan and rh9jan.



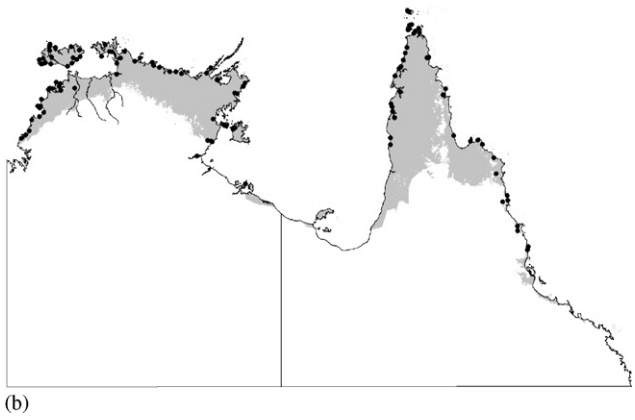


Fig. 4 – Predicted range maps for *A. farauti* ss. (a) Based on nine variables identified by jackknifing from those ranked by data mining in data set 2: coastdist, diurntemp (p2), maxtemp (p5), rangetemp (p7) rh3jan, rh3ann, rh3jul, dp3jul and dp9jul. (b) Based on all environmental variables except distance from the coast.

Seeker in data set 2 (Coastdist, p2, p5, p7, rh3jan, rh3ann and rh3jul) as well as 2 variables for atmospheric moisture ranked by KnowledgeSeeker for data set 2 (dp3jul and dp9jul). The range map for this combination (Fig. 4a) was similar to that derived from the 13 ranked variables for this species though it excluded one site (Humpty Doo) in the Northern Territory. The most significant environmental variable influencing the range of this species was distance from the coast and its removal resulted in unrealistic range predictions extending inland (Fig. 4b).

A predicted range map for *A. farauti* 2 based on eight environmental layers (Fig. 5a) included the five variables ranked by both CART and KnowledgeSeeker in data set 2 (p22, p27, dp3jul, dp9jul and rh9jul) together with 3 variables (p16, p18, and p24) ranked by KnowledgeSeeker in data set 2. This extended further south and west in the Northern Territory, further inland from the east coast of Queensland, and further south along the west coast of Cape York Peninsula than the observed distribution of this species. Several other different aggregations of variables provided similar range maps without producing one which was equivalent to the map derived from all environmental layers and the one based on the 15 ranked data mining variables.

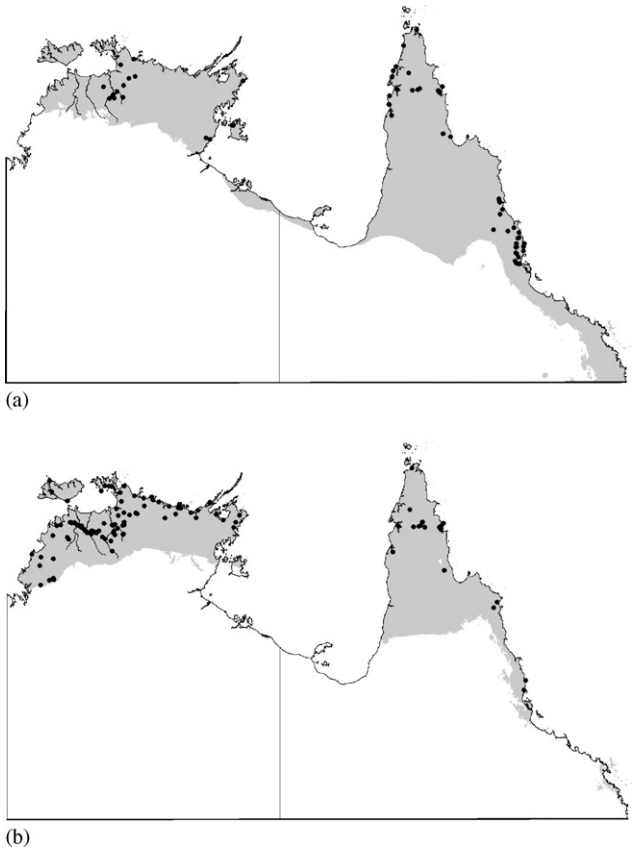


Fig. 5 – Predicted range map for *A. farauti* 2 and *A. farauti* 3 based on variables identified by jackknifing from those ranked by data mining in data set 2. (a) Range map for *A. farauti* 2 based on eight variables: rainwetqtr (p16), rainwarmqtr (p18), lowrad (p22), radwetqtr (p24), radcoldqtr (p27), dp3jul, dp9jul and rh9jul. (b) Range map for *A. farauti* 3 based on five variables: isotemp (p3), seastemp (p4), tempcoldqtr (p11), annrain (p12) and dp3jan.

The realistic range outputs for *A. farauti* 3, derived from 11 ranked variables, was maintained for a GARP run based on only four BIOCLIM parameters (p3, p4, p11 and p12) as well as one parameter for atmospheric moisture (dp3jan). This resulted in a high quality range map (Fig. 5b) which included all presence records in Queensland and the Northern Territory. The predicted range did not extend into areas in which the species has not been found, though it did include the west coast of Cape York Peninsula beyond the southern limit of record sites.

4. Discussion

4.1. *Use of data mining for ecological niche modelling*

Data mining refers to a set of computer based tools which permit exploratory data analysis to reveal patterns and relationships in databases. They include neural networks, Bayesian classification, genetic algorithms, and other machine learning procedures which have been used extensively for modelling species distributions (Guisan and Zimmerman, 2000). Classification techniques are a category of data mining methods which have been employed for this purpose. These involve decision trees and other rule-based classification tools which assign a class of the response variable (e.g. species presence or absence) to combinations of environmental predictors. This process uses computer induction to search for general principles from analysis of specific examples. Decision trees produce a prediction of class membership (e.g. species presence) by examining a learning sample which consists of data values of attributes (e.g. environmental variables) associated with that class. The tree can be expressed as decision rules which are incorporated into the model.

Classification and regression trees (CART) have been used in several studies to investigate spatial relationships between environmental predictors and species distributions. Walker (1990) employed CART and GLIM (generalised linear models) to model the realised distributions of three kangaroo species in Australia. The results showed that models derived from both methods approximated the observed distributions of the different species. CART prediction errors for eastern grey and western grey kangaroos were less than the equivalent CART errors for red kangaroos. In a subsequent contribution Walker and Cocks (1991) described the HABITAT procedure for modelling environmental envelopes which enclose species presence records. This method incorporates the CART algorithm to identify a reduced set of environmental parameters for explaining species distribution and to divide the envelope into sub-envelopes with differing degrees of membership (i.e. different proportions of record sites). The HABITAT model achieved superior results when applied to the distribution data set for red kangaroos with 5.5% prediction errors compared with 13% prediction errors for this species with the CART-only model used in the previous study.

The distribution of shrub species in the southwest ecoregion of California were predicted using generalised additive models, generalised linear models and CART by Franklin (1998). Spatially interpolated bioclimatic and topographic variables were evaluated using species presence/absence data as the dependent variable. All three methods produced models with similar accuracy for a given species though CART yielded prediction errors which were lower than the other two methods.

Decision trees have been utilised for modelling the distribution of soil properties across the landscape. Bui et al. (2006) used the Australian Soil Resources Information System database (ASRIS) to generate models of soil properties using the commercial data mining software, C5.0 for categorial responses and CUBIST for continuous responses (http://www.rulequest.com). The models were spatially evaluated and interpreted by tabulating the variables selected by the decision trees and mapping the geographical extent over which individual variables and combinations of variables were used as predictors. Summerell et al. (2000) used KnowledgeSeeker to model the distribution of parna (a wind-blown clay) in the Young district of New South Wales. Information on the geographic extents of parna deposits derived from analysis of air photos and field surveys was used as a training data set to explore spatial relationships with topographic variables. Values of the landscape variables at the position of each

parna and no-parna category were tabulated and imported into KnowledgeSeeker. The rules derived from this process were applied to the study area to predict the occurrence of parna in the broader landscape and the predictions were confirmed by ground truthing.

Termansen et al. (2006) combined two machine learning approaches: Bayesian classification and genetic algorithms to model spatial distributions of species from environmental variables. Bayesian classification was used to identify critical environmental thresholds which define species suitability. Optimisation of the environmental niche was based on a purpose-built genetic algorithm which searches for the values of environmental thresholds that maximise the fit between the observed species presence/absence patterns and the posterior probabilities from the Bayes model.

This approach was applied to modelling the distribution of native plants in the British Isles to identify the most important variables which explain species distributions. The predictive performance of the Bayes based genetic algorithm was very similar to the results obtained using generalised additive models, generalised linear models and CART.

4.2. *Application of decision trees in the present study*

Many species data sets, such as museum collections, consist of presence data without equivalent absence data. A particular advantage of our north Australian mosquito data sets is that their classification into record sites where species were found as well as no-record sites where particular species were not collected enabled statistical exploration of the data with decision tree software to identify the key environmental variables associated with presence and absence of the different mosquitoes. Unlike the studies of other workers discussed above (e.g. Franklin, 1998; Termansen et al., 2006; Summerell et al., 2000) the decision tree rules generated in CART and KnowledgeSeeker were not used to model species distributions. Also, CART was not incorporated into the modelling algorithm as was done in the HABITAT procedure (Walker and Cocks, 1991). The geographical coordinates of record and no-record sites were explicitly excluded from the analyses so that the decision tree algorithms were used to empirically rank the environmental variables without considering their spatial context. Niche modelling in this study was solely undertaken with GARP. Our application of data mining was to facilitate the statistical ranking of the environmental variables in order to highlight reduced sets of variables for inclusion as environmental parameters in GARP models. Their importance as predictors of species distributions was assessed by comparing the quality of model outputs based on ranked variables with those using the full complement of environmental information.

4.3. *Quality of survey data*

The quality of GARP range predictions were assessed by comparing them with the species distribution records from survey collections. It might be considered that our mosquito survey methods involving a different sector of northern Australia each year may provide a less reliable indication of realised species distributions than collections based on multi-year observations in the same area. Our surveys were constrained by the need to make collections when mosquito density is highest after the wet season. It would not be possible to survey the whole of this vast (500,000 km^2) and remote area during 2–3 months in a single year and repeat the collections in subsequent years. However, additional collection data are available in some key areas of species presence and absence. Van den Hurk et al. (2000) made monthly longitudinal observations of the *A. farauti* group at six study sites representing three habitat types in the Cairns area of northern Queensland between August 1995 and September 1997. The results showed that the dominant species were present at each collecting site throughout the 2-year period even though there was significant seasonal variability in their abundance. The species presence records at the different longitudinal study sites accord with those derived from our survey localities, 33 of which were within the same geographical area around Cairns. Investigations of arboviruses isolated from mosquitoes in the Gulf plains region of Queensland during April 2000 involved extensive collections around Karumba on the coast and nearby inland towns of Normanton and Croydon (Van den Hurk et al., 2002). More than 24,000 *Anopheles* were processed in this study but none were identified as *A. farauti* sl. These no-record findings are in agreement with our results for the 1987 survey which recorded a complete absence of the *A. farauti* group among 1630 *Anopheles* specimens collected from 118 localities along the southeastern coast of the Gulf of Carpentaria between Karumba and the Northern Territory border.

Surveys conducted between 1988 and 1990 indicated *A. farauti* 3 to be the most common and widely distributed species in the Northern Territory. On the other hand, *A. farauti* 2 had a restricted range in this part of Australia. We found record sites of the latter species at several isolated localities on the northern and eastern coasts of Arnhem Land but this mosquito was only collected inland from the coast in a relatively small area on the floodplain of the South Alligator River. Both *A. farauti* 2 and *A. farauti* 3 were common in this area, but *A. farauti* 3 was the only species found further west in the floodplains of the Mary and Adelaide Rivers. In April 1992 a series of seven overnight trap collections were made in three sites near Cooinda adjacent to the South Alligator River where *A. farauti* 2 was collected in the 1989 survey. All of 375 specimens of *A. farauti* sl were identified as *A. farauti* 2 (R.D. Cooper, T. Burkot, unpublished data). An additional survey was undertaken in March–April 1994 to investigate further the inland distribution of the two species in the Northern Territory. Trap collections yielded 123 *A. farauti* 3 from 23/25 localities on the Mary and Adelaide River floodplains. Other mosquitoes collected during this survey included 1436 other anophelines, none of which were identified as *A. farauti* 2. The data accumulated from these additional surveys in the Northern Territory and Queensland gave the same results as the progressive annual surveys: species found present in the annual surveys were also found in the additional surveys; and species not collected in the annual surveys were not collected in the additional surveys. This supports the view that our mosquito survey methods based on progressive surveys in different sectors each year provided a good indication of the realised distribution of the *A. farauti* group in northern Australia.

4.4. *Quality of range predictions*

The GARP series based on the data mining results of data set 2 showed that the quality of baseline range predictions with all 41 environmental layers was maintained when the inputs for each species were reduced to between 11 and 15 layers. Additional jackknifing analyses of the variables ranked by data mining led to further reduction of environmental coverages for *A. farauti* ss and *A. farauti* 3 without adversely affecting range modelling outputs. For *A. farauti* ss the minimal set of environmental layers commensurate with high quality range models included three layers for temperature (p2, p5 and p7); five layers for atmospheric moisture (rh3ann, rh3jan, rh3jul, dp3jul and dp9jul) and distance from the coast. The latter coverage is spatially related and it could be said that its inclusion among the environmental layers may artificially constrain the predictive ability of the GARP modelling system. However, we believe that there is a rationale for its addition in the models generated for *A. farauti* ss. This species is the only member of the *A. farauti* group in northern Australia which breeds in brackish water (Sweeney, 1987). Despite the fact that it is sometimes collected in freshwater larval sites, occasionally in association with *A. farauti* 2 and *A. farauti* 3 (Sweeney et al., 1990), its distribution is predominantly coastal as 120 out of the 127 record sites for this study were <10 km inland and only 1 record site (Humpty Doo) was >20 km inland. Furthermore, this species is the most common malaria vector in coastal areas throughout the South West Pacific Region. It was found within 10 km of the coast in 197 of 239 record sites in Papua New Guinea (Cooper et al., 2002). The inland collections from Papua New Guinea have been shown by DNA analysis to belong to a different genotype (and possibly a different species) to the coastal material (Beebe and Cooper, 2002). Our investigations of various combinations of climatic layers identified by data mining failed to reveal one in which coastal range was accurately predicted by GARP. On the basis of the known coastal distribution of this species we considered that the inclusion of a layer for distance from the coast was justified as a surrogate factor for unidentified biotic or abiotic factor(s) which restrict the range of this species.

Subsequent to the present study, we continued our niche modelling of *A. farauti* ss with the use of the boundary *U*-test, a recently developed software tool that permits analysis of environmental gradients across distributional boundaries (Bauer and Peterson, 2005). In this procedure, the Mann–Whitney *U*-test was used to contrast values of environmental layers inside and outside a boundary line which extended 5 km around the coast of northern Australia. This approach identified the same key variables which were identified by data mining as being associated with the range of *A. farauti* ss but it also highlighted the importance of elevation which was not among the variables ranked by KnowledgeSeeker and CART (Sweeney et al., 2006). Almost 90% of the record sites of *A. farauti* ss were within 5 km from the coast whereas most of the no-record sites were more than 50 km inland. It appears that the lack of sufficient data points on either side of the narrow range limit of this species impaired the data mining software from accurately ranking variables such as elevation which have a steep gradient across the boundary.

Models based on all environmental layers, as well as those derived from the key variables identified by data mining, predicted occurrence for all three species along the south western side of Cape York Peninsula to 17°S near Karumba in the south west corner of the Gulf of Carpentaria. These predictions do not accord with our collection data for this part of northern Australia. *A farauti ss* was not found beyond 14.3°S, the southernmost record sites of *A. farauti* 2 and *A. farauti* 3 were 13.5°S, and there were 42 no-record sites between these localities and Karumba. However, our survey results do not precisely define the species distributions for this region as Kay (1985) collected *A. farauti* sl at Kowanyama (15.5°S), though it was not common, with only 25 specimens among >6000 mosquitoes collected during 1974–1975. Nevertheless the observed presence and absence data of our surveys strongly suggest that these mosquitoes do not extend as far as the predicted range limits of the GARP models. In this remote area of northern Australia there are only 3 meteorological stations for temperature (at Weipa, Kowanyama and Karumba, see Fig. 1a) which contributed to the climate surfaces (Hutchinson and Kesteven, 1998). Thus, it is possible that there is insufficient meteorological data to support adequate interpolation of the climatic variables in this region and this may explain the localised deficiencies in the niche models.

There are many reports on the importance of humidity for the survival of adult mosquitoes in the laboratory and in the field (reviewed by Clements (1963)). Diurnal and seasonal variability of this factor was considered by including dewpoint and relative humidity layers for the daily times (9 a.m. and 3 p.m.) at which this data is collected at weather stations during the months coinciding with mid-wet season (January) and mid-dry season (July). Previous studies of the distribution and prevalence of *A. farauti* sl in the Cairns area of northern Queensland during World War 2 indicated that these mosquitoes flourish during the warm and humid conditions of the north Australian wet season but the adults retreat to sheltered positions in the bush during the dry season (Roberts, 1948). However, Van den Hurk et al. (2000) found that the numbers of *A. farauti* sl collected in traps showed no correlation with relative humidity recorded on the night of trapping. These authors concluded that this may have been due to the low variability in relative humidity values in their study areas. In the present study atmospheric moisture variables were shown to be among the key indicator variables associated with range of all members of the *A. farauti* group in northern Australia ranging from 5 of 8 variables for *A. farauti* ss (rh3jan, rh3ann, rh3jul, dp3jul and dp9jul), 4 of 15 variables for *A. farauti* 2 (dp3jul, dp9jan, dp9jul, and rh9jul) and 1 of 5 variables for *A. farauti* 3 (dp3jan). It might be argued that there is a degree of intercorellation between the atmospheric moisture variables. However, omission of those identified by data mining resulted in models of inferior quality. The need for the inclusion of several atmospheric moisture coverages to obtain good predictive range models of *A. farauti* ss may reflect the biological requirements of this coastal species as proximity to the sea is associated with higher aerial moisture conditions than the arid interior of the continent.

The final GARP models for *A. farauti* 3, derived from only five environmental layers, resulted in a high quality range prediction which was similar to that based on all 41 environmental

variables. It included all record sites without extending to the no-record collection areas for this species in the southern Gulf of Carpentaria and inland in the Northern Territory beyond 14.8°S. *A farauti* 3 is the species of the *A. farauti* group which is most adapted to arid conditions. Its distribution extends further into the arid inland of the Northern Territory than *A. farauti* ss and *A. farauti* 2 and, unlike the other two species, it is confined to the drier south western area of Papua New Guinea and is not found in the wet tropical regions of the country (Cooper et al., 2002). The association of only one atmospheric moisture factor with the best range models for this species may be indicative of its ability to tolerate more arid conditions than the other members of the group.

The range models for *A. farauti* 2 based on all environmental layers and also the 15 ranked layers identified by data mining in data set 2 were only of fair quality. They encompassed all of the survey record sites but they also included the southern coast of the Gulf of Carpentaria and the north eastern and northwestern part of the Northern Territory, from the Mary River to the coast west of Darwin, and the interior of Arnhem Land where this species was not collected. Moreover, further reduction in the number of environmental variables using jackknifing produced GARP range models of poor quality. Also, unlike the other two species, there was no increase in χ^2 values of the KnowledgeSeeker ranked variables for *A. farauti* 2 in data set 2 compared to those in data set 1. These results might imply that the environmental controls on the range of this species are more diffuse than those influencing the ranges of the other members of the *A. farauti* group in Australia. Recent observations using molecular-based techniques have shown that the population of *A. farauti* 2 in the Northern Territory is genetically different to the population of this species in Queensland (N.W. Beebe, unpublished data). If these populations differ in their biological and environmental characteristics it may confound ecological niche modelling investigations and may explain our inability to obtain high quality range models for this species in northern Australia. The fact that there are no meteorological stations in the interior of Arnhem Land may also have contributed to weakness in the interpolated climate layers and resulted in overprediction of *A. farauti* 2 in this region.

There is no doubt that some of the no-record collections of the different species were not true absence records as they were within the species observed distribution limits. Nevertheless, their inclusion in data set 1 permitted the identification of significant variables among the environmental layers which were associated with species presence. These results suggest that this software is sufficiently robust to function effectively even when some of the records in the dependent variable are incorrectly classified. The exclusion of no-record sites within the range predicted by GARP for each species in data set 2 led to an increase in the statistical significance of the ranked variables as well as an increase in the numbers of common variables ranked by both data mining procedures. CART outputs are not based solely on variables in primary splits of the data set as the procedure calculates the contribution of other variables (Steinberg and Colla, 1995) whereas KnowledgeSeeker uses chi-squares to investigate each variable separately. It is interesting that our results using KnowledgeSeeker, which does not consider the role of surrogates in the tree-splitting process, did not appear inferior to those achieved with CART.

5. Conclusions

An extensive biological data set of mosquito survey records collected over 10 years in the remote north of Australia has been successfully analysed. The consistent agreement of follow up surveys with the original data set is a strong argument for the quality of the collections and for the quality of the analyses which were derived from them. The similar performance of CART and KnowledgeSeeker in determining the important environmental variables associated with species presence and absence is also an indicator of the worth of these analyses.

Atmospheric moisture is a key predictor for all three species of the *A. punctulatus* group in Australia. The environmental parameters which define the realised distributions of *A. farauti* ss and *A. farauti* 3 were well described by a combination of ecological niche modelling with GARP and the ranking procedures of decision trees. This approach was less satisfactory for explaining the environmental factors associated with the distribution of *A. farauti* 2 in the continent. Insufficient meteorological stations in some isolated parts of the survey area may have been responsible for localised weaknesses in the interpolated climate variables. This might explain over predictions derived from the niche models for all three species along the southwest side of Cape York Peninsula and for *A. farauti* 2 in the Northern Territory.

Acknowledgements

This work was supported by grant 211608 from the Australian National Health and Medical Research Council. We thank Mike Hutchinson for discussions and guidance in using ANUSPLIN. Thanks are also due to John Stein for generating a 0.01° Digital Elevation Model of northern Australian using ANUDEM 5.1. We also thank an anonymous reviewer who made helpful comments on an earlier version of this manuscript and a second anonymous reviewer who offered most useful suggestions for the initial and present versions of the manuscript.

REFERENCES

Anderson, R.P., Lew, D., Peterson, A.T., 2003. Evaluating predictive models of species' distributions: criteria for selecting optimal models. Ecol. Model. 162, 211–232.

Bauer, J.T., Peterson, A.T., 2005. Visualizing environmental correlates of species geographical range limits. Diversity Distrib. 11, 275–278.

Beebe, N.W., Cooper, R.D., 2002. Distribution and evolution of the *Anopheles punctulatus* group (Diptera: Culicidae) in Australia and Papua New Guinea. Int. J. Parasitol. 32, 563–574.

Biggs, D., de Ville, B., Suen, E., 1991. A method for choosing multiway partitions for classification and regression trees. J. Appl. Statsist. 18, 49–61.

Breiman, L., Friedman, J., Olshen, R., Stone, C., 1984. Classification and Regression Trees. Pacific Grove, Wadsworth.

Bui, E.N., Henderson, B.L., Viergever, K., 2006. Knowledge discovery from models of soil properties developed through data mining. Ecol. Model. 191, 431–446.

Clements, A.N., 1963. The Physiology of Mosquitoes. Pergamon Press, Oxford.
Cooper, L., Cooper, R.D., Burkot, T.R., 1991. The *Anopheles punctulatus* complex: DNA probes for identifying the Australian species using isotopic, chromogenic and chemiluminescence detection systems. Exp. Parasitol. 73, 27–35.
Cooper, R.D., Frances, S.P., Sweeney, A.W., 1995. Distribution of members of the *Anopheles farauti* complex in the Northern Territory of Australia. J. Am. Mosq. Contr Assoc. 11, 66–71.
Cooper, R.D., Frances, S.P., Waterson, D.G.E., Piper, R.G., Sweeney, A.W., 1996. Distribution of anopheline mosquitoes in northern Australia. J. Am. Mosq. Contr. Assoc. 21, 656–663.
Cooper, R.D., Waterson, D.G.E., Frances, S.P., Beebe, N.W., Sweeney, A.W., 2002. Speciation and distribution of the *Anopheles punctulatus* (Diptera: Culicidae) Group in Papua New Guinea. J. Med. Entomol. 39, 16–27.
Guisan, A., Zimmerman, N.E., 2000. Predictive habitat distribution models in ecology. Ecol. Model. 135, 147–186.
Franklin, J., 1998. Predicting the distribution of shrub species in southern California from climate and terrain-Derived variables. J. Veg. Sci., 733–748.
Houlder, D., Hutchinson, M., Nix, H., McMahon, J., 1999. ANUCLIM Version 5 User Guide. Centre for Resource and Environmental Studies, Canberra.
Hutchinson, M., 1997. ANUDEM Version 4.6. Centre for Resource and Environmental Studies, Australian National University, Canberra.
Hutchinson, M., 2003. ANUSPLIN Version 4.2 User Guide. Centre for Resource and Environmental Studies, Australian National University, Canberra.
Hutchinson, M., Kesteven, 1998. Monthly Mean Climate Surfaces for Australia. Centre for Resource and Environmental Studies, Australian National University, Canberra, Publically available at http://cres.anu.edu.au/outputs/climatesurfaces.
Kay, B.H., 1985. Man-mosquito contact at Kowanyama Northern Queensland, Australia. J. Am. Mosq. Contr. Assoc. 1, 191–194.
Levine, R.S., Peterson, A.T., Benedict, M.Q., 2004a. Distribution of members of *Anopheles quadrimaculatus* Say s.l. (Diptera: Culicidae) and implications for their roles in malaria transmission in the United States. J. Med. Entomol. 41, 607–613.
Levine, R.S., Peterson, A.T., Benedict, M.Q., 2004b. Geographic and ecologic distributions of the *Anopheles gambiae* complex predicted using a genetic algorithm. Am. J. Trop. Med. Hyg. 70, 105–109.
Mahon, R.J., 1984. The status and means of identifying the member of the *Anopheles farauti* Laveran complex of sibling species. In: Malaria. Proceedings of a Conference to Honour Robert H. Black. Australian Government Publishing Service, Canberra, pp. 152–156.
Peterson, A.T., Cohoon, K.P., 1999. Sensitivity of distributional prediction algorithms to geographic data completeness. Ecol. Model. 117, 159–164.
Roberts, F.H.S., 1948. The distribution and seasonal prevalence of anopheline mosquitoes in north Queensland. Proc. R. Soc. Qld. 59, 93–100.
Rohe, D.L., Fall, R.P., 1979. A miniature battery powered CO_2 baited light trap for mosquito borne encephalitis surveillance. Bull. Soc. Vect. Ecol. 4, 24–27.
Scachetti-Pereira, R., 2003. DesktopGarp. Publically available at http://www.lifemapper.org/desktopgarp.
Steinberg, D., Colla, P., 1995. CART: Tree-Structured Non-Parametric Data Analysis. Salford Systems, San Diego.
Stockwell, D., Peters, D., 1999. The GARP modelling system: problems and solutions to automated spatial prediction. Int. J. Geogr. Inf. Sci. 13, 143–158.
Summerell, G.K., Dowling, T.I., Richardson, D.P., Crawford, J.R., 2000. Modelling current parna distribution in a local area. Aust. J. Soil Res. 34, 867–878.
Sweeney, A.W., 1987. Larval salinity tolerances of the sibling species of *Anopheles farauti*. J. Am. Mosq. Contr Assoc. 3, 589–592.
Sweeney, A.W., Beebe, N.W., Cooper, R.D., Bauer, J.T., Peterson, A.T., 2006. Environmental factors associated with distribution and range limits of malaria vector *Anopheles farauti* in Australia. J. Med. Entomol. 43, 1068–1075.
Sweeney, A.W., Cooper, R.D., Frances, S.P., 1990. Distribution of the sibling species of *Anopheles farauti* in the Cape York Peninsula, northern Queensland, Australia. J. Am. Mosq. Contr. Assoc. 6, 425–429.
Termansen, M., McClean, C.J., Preston, C.D, 2006. The use of genetic algorithms and Bayesian classification to model species distributions. Ecol. Model. 192, 410–424.
Van den Hurk, A.F., Cooper, R.D., Beebe, N.W., Williams, G.M., Bryan, J.H., Ritchie, S.A., 2000. Seasonal abundance of *Anopheles farauti* (Diptera: Culicidae) sibling species in far north Queensland, Australia. J. Med. Entomol. 37, 153–161.
Van den Hurk, A.F., Nisbett, D.J., Foley, P.N., Ritchie, S.A., Mackenzie, J.S., Beebe, N.W., 2002. Isolation of arboviruses from mosquitoes (Diptera: Culicidae) collected from the gulf plains region of northwest Queensland, Australia. J. Med. Entomol. 39, 786–792.
Walker, P.A., 1990. Modelling wildlife distributions using a geographic information system: kangaroos in relation to climate. J. Biogeogr. 17, 279–289.
Walker, P.A., Cocks, K.D., 1991. Habitat: a procedure for modelling a disjoint environmental envelope for a plant or animal species. Glob. Ecol. Biogeogr. Lett. 1, 108–118.