Author: R. Whitsed, R. Corner and S. Cook

Title: A model to predict ordinal suitability using sparse and uncertain data

Journal: Applied Geography **ISSN**: 0143-6228

Year: 2012 **Volume:** 32 **Issue:** 2 **Pages:** 401-408

Abstract: We describe the development of the algorithms that comprise the Spatial Decision Support System (SDSS) CaNaSTA (Crop Niche Selection in Tropical Agriculture). The system was designed to assist farmers and agricultural advisors in the tropics to make crop suitability decisions. These decisions are frequently made in highly diverse biophysical and socioeconomic environments and must often rely on sparse datasets.The field trial datasets that provide a knowledge base for SDSS such as this are characterised by ordinal response variables. Our approach has been to apply Bayes' formula as a prediction model.This paper does not describe the entire CaNaSTA system, but rather concentrates on the algorithm of the central prediction model. The algorithm is tested using a simulated dataset to compare results with ordinal regression, and to test the stability of the model with increasingly sparse calibration data. For all but the richest input datasets it outperforms ordinal regression, as determined using Cohen's weighted kappa. The model also performs well with sparse datasets. Whilst this is not as conclusive as testing with real world data, the results are encouraging.

URLs: http://researchoutput.csu.edu.au/R/-?func=dbin-jump-full&object_id=29949&local_base=GEN01-CSU01; http://dx.doi.org/10.1016/j.apgeog.2011.06.016

Author Address: rwhitsed@csu.edu.au

CRO Number: 29949

A model to predict ordinal suitability using sparse and uncertain data

R. Whitsed [a], R. Corner [b] and S. Cook [c]

[a] Institute for Land, Water and Society, Charles Sturt University, PO Box 789, Albury NSW 2640, Australia. Email rwhitsed@csu.edu.au. Phone +61 2 6051 9641. Fax +61 2 6051 9897. *Corresponding author.*

[b] Department of Spatial Sciences, Curtin University of Technology, GPO Box U1987, Perth WA 6845, Australia. Email r.corner@curtin.edu.au.

[c] CIAT (International Center of Tropical Agriculture), AA 6713, Cali, Colombia. Email s.cook@cgiar.org.

Abstract

We describe the development of the algorithms that comprise the Spatial Decision Support System (SDSS) CaNaSTA (Crop Niche Selection in Tropical Agriculture). The system was designed to assist farmers and agricultural advisors in the tropics to make crop suitability decisions. These decisions are frequently made in highly diverse biophysical and socioeconomic environments and must often rely on sparse datasets.

The field trial datasets that provide a knowledge base for SDSS such as this are characterised by ordinal response variables. Our approach has been to apply Bayes' formula as a prediction model.

This paper does not describe the entire CaNaSTA system, but rather concentrates on the algorithm of the central prediction model. The algorithm is tested using a simulated dataset to compare results with ordinal regression, and to test the stability of the model with increasingly sparse calibration data. For all but the richest input datasets it outperforms ordinal regression, as determined using Cohen's weighted kappa. The model also performs well with sparse datasets. Whilst this is not as conclusive as testing with real world data, the results are encouraging.

Keywords: spatial modelling; Bayesian probability modelling; CaNaSTA; sparse data; agriculture

1. Introduction

1.1. Overview

This paper describes the results of research undertaken to develop a model to be embedded in a spatial decision support system (SDSS) to predict suitability of crops in the tropics and subtropics. The aim was to identify a suitable model that could use typical trial database data to predict suitability of various crops for different locations. An SDSS, CaNaSTA (Crop Niche Selection in Tropical Agriculture), was developed using Bayesian probability modelling. A description is then given of the algorithms employed in CaNaSTA, and the validity of the algorithms is assessed using a simulated case study. Other papers (Läderach et al., 2006; Atzmanstorfer et al., 2007) describe real world applications of the SDSS, and further research is planned to verify CaNaSTA in real world situations.

In the database that we initially employed (RIEPT database, Barco et al., 2002; see section 2.1), and in many similar databases, the suitability of crops at different trial sites is recorded ordinally, i.e. at discrete levels of suitability. In addition, trial databases often record trials in a narrow range of environmental conditions for many species. Therefore a model was needed that could handle ordinal response data and very small calibration datasets.

1.2. Bayesian approach

The approach used in this research is Bayesian probability modelling (Pearl, 1990; Jensen, 1996). This approach satisfies the need for ordinal responses and allows the incorporation of both trial data and expert knowledge as inputs. In addition, uncertainty can easily be described and quantified using Bayesian equations. Bayesian probability modelling is described below in Section 2.2.

Bayes' theorem has been described and implemented in GIS applications for at least two decades, with early examples including Skidmore (1989), Fischer (1990), Aspinall (1992), Aspinall and Veitch (1993) and Bzreziecki et al. (1993). This promising start was largely abandoned in favour of more complex statistical models, but more recent descriptions of the Bayesian approach can be found in Corner et al. (2002), Zhou et al. (2004) and La Morgia et al. (2008), using approaches based on those used by Skidmore (1989) and Aspinall (1992). La Morgia et al. (2008) also discuss the inclusion of expert knowledge and suggest that the Bayesian approach in habitat suitability modelling is most useful when dealing with rare species or low detectabiltiy.

This research analyses the performance of the Bayesian algorithm with increasingly smaller calibration datasets, and shows that under some conditions, the algorithm performs well with very little input data. We also compare the Bayesian algorithm with ordinal regression (McCullagh and Nelder, 1989), and show that with our case study, our algorithm outperforms ordinal regression with sparse calibration datasets.

1.3. Uncertainty in agriculture

Agricultural decisions often have to be made based on very little information. For example, detailed trial data on new crop species will rarely be available for a farmer's own biophysical and management environment. Farmers are generally only interested in crop species that will thrive in their particular location or niche. Farmers in tropical environments often have few resources and high levels of uncertainty in their agricultural environment. The uncertainties that farmers face include environmental variability (both spatial and temporal), ignorance (lack of information) and uncertainty surrounding the results of models designed to assist the decision-making process.

The SDSS is designed to help farmers in these environments select which crop to plant where under conditions of uncertainty and risk. In particular we are addressing the impact of sparseness and variability in predictor variables in the spatial model.

1.4. Outline of paper

This paper begins by briefly describing the RIEPT database to characterise the kind of data that the model needs to work with. Bayesian probability modelling is then described, and in

particular the equation used in CaNaSTA to predict probability distributions for locations based on the set of environmental conditions at each location. A case study using simulated data is presented, examining how the algorithm performs with successively smaller input dataset, and comparing the algorithm with ordinal regression. This is followed by the results of the case study. The paper concludes with discussion and conclusion sections.

2. Data and methods

2.1. RIEPT database and expert knowledge

2.1.1. RIEPT database

The RIEPT[1] database (Barco et al., 2002), held by the International Center for Tropical Agriculture (CIAT) contains adaptation data for forage trials mainly throughout Central and South America, spanning 1979-1992, for the purpose of evaluating grasses and legumes in locations representative of major tropical ecosystems. Variables describing the trial sites include elevation, climate and soil data. Response variables include adaptation, percentage cover, dry matter weight and number of plants.

It should be noted that variables describing the trial sites can in most cases also be derived from GIS (Geographical Information System) layers, assuming the location of the trial site is properly recorded. The database obviously only records variables for locations where trial sites are situated, therefore in order to predict suitability for other locations, GIS layers must be used.

2.1.2. Expert knowledge

A good deal of expert knowledge exists on which forages and other crops are suitable in which locations. This expert knowledge can take many forms and can sometimes be valuable for filling in gaps in the data available in a database. It is therefore useful to have a model which allows for the incorporation of some expert knowledge, where available.

2.1.3. Simulated data

Precisely because of the sparseness of data in the RIEPT database, they cannot be used for validating the model – there are simply not enough trials under different circumstances for the various forage species. Therefore in order to validate the model, a simulated dataset was created, mimicking typical patterns seen for forage species in the RIEPT database. This simulated dataset is described below in section 2.3.

[1] *Red Internacional de Evaluación de Pastos Tropicales* (International Network for the Evaluation of Tropical Pastures)

2.2. Bayesian probability modelling

Bayesian methods provide a 'formalism for reasoning under conditions of uncertainty, with degrees of belief coded as numerical parameters, which are then combined according to rules of probability theory' (Pearl, 1990). A Bayesian model defines prior and conditional probability distributions and then uses these to calculate posterior probability distributions. The probability distribution may be derived from data, set by experts or defined from a combination of data and expert opinion.

Bayesian methods have been applied to species distribution (Aspinall, 1992; Aspinall and Veitch 1993) and relative abundance of species (Gorman et al., 2007). Stassopoulou et al. (1998) used a Bayes' network with a GIS in order to combine information from different sources of data to classify risk of desertification in forest areas. Grêt-Regamey and Straub (2006) used Bayesian networks to classify avalanche risk in a GIS, and Aalders (2008) applied Bayesian methods to model land-use decision behaviour. Asadi and Hale (2001) used Bayesian methods and GIS to predict mineral deposits. Corner et al. (2002) employed similar methods to map soil attributes. Our approach most closely mirrors that of Asadi and Hale (2001) and Corner et al. (2002).

A prior probability is an initial estimate that may be modified once more information becomes available. Prior probability that Y is in state y_i is denoted $P(Y = y_i)$, which for simplicity can be written $P(y_i)$. Joint probability refers to the probability of two events occurring together, and is denoted by $P(X = x_i, Y = y_j)$ or $P(x_i, y_j)$. Conditional probability is the probability of Y being in state y_j, given that X is in state x_i, and is denoted $P(Y = y_j | X = x_i)$ or $P(y_j | x_i)$. Posterior probability is the conditional probability of Y given multiple sets of evidence $X_1, X_2, \ldots X_n$, denoted $P(Y = y_i | X_1 = x_{i1}, X_2 = x_{i2}, \ldots, X_n = x_{in})$ or $P(y_j | x_{i1}, x_{i2}, \ldots x_{in})$.

Based on Bayes' inversion formula (Pearl, 1990), the following equation can be derived for conditional probability:

$$P(y_j \mid x_i) = \frac{P(y_j)P(x_i \mid y_j)}{\sum_{j=1}^{n} P(y_j)P(x_i \mid y_j)} \quad (1)$$

where $P(y \mid x)$ denotes conditional probability and $P(y)$ denotes prior probability as described above.

Equation 1 can be used to derive an equation for the posterior probability distribution of a response variable Y given a number of predictor variables $X_1, X_2, \ldots X_n$. For example, the response variable could be adaptation, with predictor variables elevation, rainfall and temperature.

Suppose there are *n* conditionally independent predictor variables denoted X_1, X_2, … X_n. Then the following posterior probability equation can be derived:

$$P(y_j \mid x_{i1}, x_{i2}, ..., x_{in}) \propto P(y_j) \prod_{k=1}^{n} \left(\frac{P(y_j \mid x_{ik})}{P(y_j)} \right) \quad (2)$$

This is the equation used in CaNaSTA to predict posterior probability distributions for locations based on the set of environmental conditions at each location. Defining the model in terms of probabilities also gives some measure of uncertainty. The probability distribution of the response at a given location can be interpreted in terms of the certainty that the true response value is in a given state.

This equation is adjusted slightly in cases where prior or conditional probability values are zero, as will often happen with sparse datasets. If $P(y_j) = 0$ for any value of j, then Equation 2 will fail when attempting to divide by zero. The approach is to replace the prior probability values with a reasonable non-zero estimate. This situation can occur when there are no entries in the database of trials of a species where the response is a particular value (e.g. adaptation = poor). In this case the approach is to set $P(y_j) = \alpha$, where α is a very small positive number. Joint probabilities are set to α/j. This specifies an equal distribution which implies, correctly, that no information is known about the distribution (complete ignorance).

Where conditional probabilities are zero, these are again adjusted to a small positive number, and the non-zero conditional probabilities are then normalised. If all conditional probabilities are zero across a predictor variable (i.e. there are no entries in the database of trials of a species where the predictor variable is in a specific class), then prior probabilities of the response variable are used.

Equation 2 makes the assumption that the predictor variables are conditionally independent. Conditional independence (CI) means that knowing the state of one variable has no bearing on the probability distribution of another variable once the state of the response variable is known. As Aspinall (1992) points out, the requirement of CI is often not met when dealing with environmental data.

The likelihood of datasets being conditionally dependent can however be lessened by using fewer datasets. In addition, some authors (Aspinall, 1992; Corner et al., 2002) question whether CI is operationally important. Corner et al. (2002) suggest that functional independence is more critical. Even if two datasets are statistically dependent, if they have different meanings in the model then it may still be valid to include them both.

2.3. Case study

The implementation of Equation 2 is illustrated and tested using a simulated dataset of trial locations in South America. Three predictor variables, elevation, rainfall and length of dry season, are used to predict the response variable adaptation at these locations. We compare the Bayesian model with ordinal regression to test the overall performance of the Bayesian model. We chose ordinal regression as a model suited to ordinal responses in classification. Ordinal regression models are a variation of generalised linear models (GLM), modified specifically to deal with ordinal response variables (McCullagh and Nelder, 1989; Guisan and Harrell, 2000). We used ordinal regression as implemented in SPSS using the methodology of McCullagh and Nelder (1989).

We also test how the model performs with progressively smaller calibration sets. Small calibration sets are typical of the type of decision problem we are discussing here, i.e. only a small number of trial results may be known.

2.3.1. Locations

In this case study, a number of pseudo-random locations were selected – this number was set at 256 as being a sufficiently large number, and also iteratively divisible by two in order to partition the set into increasingly smaller sets for validation. These 256 locations were selected in tropical and sub-tropical Latin America, between latitudes 34N and 34S; and longitudes 103W and 34W. These boundaries were chosen as broadly encompassing the tropics and subtropics, with the western and eastern boundaries containing all of South and Central America. Initially, a grid of evenly distributed points was created within these bounds, with a separation of two minutes between each point both in latitude and longitude. Starting with the first point and moving eastwards across each line of points, points were selected using a random number generator until approximately 1000 points remained. Finally, all points not on land were deleted, leaving 363 points, of which 256 points were then randomly selected (Figure 1).

2.3.2. Predictor variables

Any number of predictor variables can be chosen, but for the purpose of testing how well the algorithm performs with sparse data, three predictor variables have been chosen, in consultation with forage experts at CIAT. The selected variables are elevation (m), annual rainfall (mm) and length of dry season (months) (sourced from the WORLDCLIM database, Hijmans et al., 2005). Length of dry season is defined as maximum number of consecutive months with rainfall less than 60mm per month (Bonan, 2002). Elevation was chosen instead of temperature, as the two variables are highly correlated in the tropics and subtropics, but elevation data is available at a higher spatial resolution. A single measure of rainfall was chosen since, although monthly rainfall data are available, the use of 12 separate variables

on rainfall would clearly violate conditional independence. Both annual rainfall and length of dry season were derived from monthly rainfall data, and although these two variables are somewhat correlated, including length of dry season captures some seasonal variation not represented with annual rainfall. Correlation does not automatically imply that conditional independence (CI) has been violated. Testing for CI examines whether the calculated (expected) values for joint probabilities are close to what would be observed if the full conditional probability table for all variables were known. When CI cannot be empirically tested, as is generally the case, joint information uncertainty (JIU) (Press et al., 1986) is a useful measure. When JIU is close to 1, the variables are correlated and the assumption of CI is likely to be violated. When JIU is close to 0, the variables are uncorrelated and the assumption of CI is likely to hold. In our study area, JIU is low for all combinations of variables (Table 1). Therefore the assumption of CI is unlikely to be violated in this case study.

The variables were discretised for use in the algorithm in consultation with forage experts. Each variable was split into five classes, with the breaks for elevation at 500m, 1000m, 1500m and 2000m and for annual rainfall at 500mm, 800mm, 1200mm and 1800mm. For length of dry season the groupings were 0-2 months, 3-4 months, 5-6 months, 7-8 months and > 9 months.

2.3.3.Response variable

In a trial database, response variables would be available for each trial site to act as a calibration dataset. However in this simulated dataset, we needed to assign a response value to each location so that a relationship would be evident between predictor variables and response variable, similar to what might be observed in the real world. The response variable used was adaptation, with ordinal values ‘p’ (‘poor adaptation’), ‘a’ (‘average adaptation’), ‘g’ (‘good adaptation’) or ‘e’ (‘excellent adaptation’).

In order to assign these response values, transformations of each predictor variable were chosen to reproduce potentially authentic interactions between a species and the environment, based on patterns observed in forage species in the RIEPT database described in section 2.1.1. These values were assigned by transforming each of the three predictor variables to values between 0.5 and 4.4 (so that after rounding they would be between 1 and 4, where 1 represents poor adaptation and 4 represents excellent adaptation), and taking the average of the three transformations, rounding to the nearest whole number. The transformations used were a linear transformation for elevation and length of dry season and a second order polynomial approximation for rainfall. Figure 2 shows the response values for the 256 locations produced by these transformations. Trend lines are also displayed, including their 95% credible interval. These distributions were

estimated using Markov Chain Monte Carlo (MCMC) in WinBUGS (Spiegelhalter et al., 2005). The predictions have been truncated with an upper limit of 4 and a lower limit of 1.

To illustrate the feasibility of the response values, they have been compared with actual trial data from the RIEPT database for two species, *Pueraria phaseoloides* and *Zornia* (various species) (Barco et al., 2002). Using the same methodology as above, the data were fitted to regression equations (Figure 3). Each forage species will show a different pattern, but these two examples show that the simulated response values are reasonable.

These response values were used to calibrate and validate the model.

2.3.4. Equation

Equation 2 was applied to the predictor and response variable values calculated for the 256 sites. In this case there are three predictor variables, *elev* (elevation), *rain* (annual rainfall) and *dry* (length of dry season), with five classes for each variable, as described in section 2.3.2. So for example, the probability that adaptation is excellent $P(y_e)$ given elevation is below 500m (class 1), annual rainfall is between 1200 and 1800mm (class 4) and length of dry season is 3-4 months (class 2) is given by:

$$P(y_e \mid elev_5, rain_2, dry_4) \propto P(y_e)\left(\frac{P(y_e \mid elev_5)}{P(y_e)}\right)\left(\frac{P(y_e \mid rain_2)}{P(y_e)}\right)\left(\frac{P(y_e \mid dry_4)}{P(y_e)}\right) \quad (3)$$

Using the same equation to calculate the probability of good, average and poor adaptation under the same circumstances allows the equation to be normalised (i.e. the sum of the probabilities equals 1).

Therefore in order to derive the posterior probability distribution of the response variable, it is sufficient to calculate the prior probability distribution of the response variable, and the conditional probability of the response variable, conditioned on each possible state of each individual predictor variable.

2.3.5. Example data

Equation 3 is illustrated with a numerical example. Say we have 12 trial sites for a species, with elevation, annual rainfall and length of dry season for each location, as well as an adaptation response value. These data can then be recoded to comply with the breakpoints given in section 2.3.2 (Table 2).

In order to apply Equation 3, we first need to calculate prior and conditional probabilities. The prior probability distribution for adaptation can be derived directly from counts in the last column of Table 2, giving $P(y_e, y_g, y_a, y_p)$ = (7/12, 2/12, 2/12, 1/12) = (0.583, 0.167, 0.167, 0.083). This means that before we apply any data to the equation, we would expect that in

the majority of locations, the adaptation of this species would be excellent. The prior probabilities do not need adjusting in this case, as they are all non-zero.

We now need to calculate the conditional probabilities $P(Y \mid elev_i)$, $P(Y \mid rain_i)$ and $P(Y \mid dry_i)$ for i = 1..5. These results are again derived directly from counts in Table 2. These conditional probabilities are adjusted as described at the end of section 2.2. The adjusted values are shown in brackets (Table 3). Zero frequencies are substituted with 0.01, with the remaining frequencies in that class adjusted to sum to 1. Where all four frequencies in a class are zero (e.g. Elevation class 4 and Dry season class 2), they are substituted with prior probabilities.

Substituting these values into Equation 3 gives:

$P(y_e \mid elev_5, rain_2, dry_4) \propto 0.583(0.01/0.583)(0.01/0.583)(0.01/0.583) = 0.000003$

Similarly it can be calculated that:

$P(y_g \mid elev_5, rain_2, dry_4) \propto 0.12$
$P(y_a \mid elev_5, rain_2, dry_4) \propto 2.82$
$P(y_p \mid elev_5, rain_2, dry_4) \propto 0.01$

Normalising over the adaptation values gives the posterior probability distribution (0.000001, 0.04, 0.96, 0.002). Therefore there is a very high probability that adaptation under these circumstances would be average. This agrees with the single trial in the data with these conditions (Table 2). Had we not adjusted zero values to small non-zero values the result would have been (0, 0, 1, 0).

Now consider the case where elevation is in class 3, rainfall in class 4 and dry season in class 1. This scenario is not represented in the data in Table 2. Applying the same equation gives the posterior probability distribution (0.88, 0.06, 0.06, 0.005). Therefore there is a high probability that adaptation would be excellent under these circumstances, but there is some possibility it could also be good or average (i.e. there is a degree of uncertainty). Note that not adjusting zero values here would result in the distribution (0, 0, 0, 0), which cannot be normalised and is unhelpful.

2.3.6. Calibration, validation and sparseness

The set of 256 points was randomly split into two to form a calibration and a validation set, each with 128 points. This process was repeated to produce 10 different combinations of calibration and validation data sets. The calibration dataset was then used as input into Equation 3 and the resulting response values were compared with the validation set using Cohen's weighted kappa (κ_w) with quadratic weights (Cohen, 1960; Fleiss, 1981). The same calibration and validation sets were used in the ordinal regression model.

In order to test the model's response to sparseness, the calibration set was progressively and randomly halved to produce calibration sets of 64, 32, 16, eight and four points. These calibration sets were again used to predict the validation data in both Equation 3 and ordinal regression.

Additional runs were carried out for sets of eight and four points, to test the effect of the range of response values in the calibration set. The sets of four and eight were split into two divisions each – 'high' (H) where the calibration set contains three or four out of four distinct response values, and 'low' (L) where the calibration set contains only one or two out of four distinct response values. The model's performance was then compared between the 'high' and 'low' calibration sets.

3. Results

Figure 4 summarises the results of the ten runs each carried out on calibration sets of 128, 64, 32, 16, eight and four points, using both the current model and ordinal regression.

κ_w was calculated for the predicted class, compared with the actual class in the validation set. The mean values of κ_w are slightly higher for ordinal regression with the largest calibration sets, but higher for the current model with smaller calibration sets. With the current model, values of κ_w remain high (means ranging from 0.84 (128 points) to 0.79 (16 points)) for larger number of points. Values drop slightly (mean = 0.74) for eight points and substantially (mean = 0.54) for four points. The range of κ_w also increases as the set size decreases. With four points, two of the runs returned $\kappa_w = 0$, signifying agreement no greater than chance. This is because, for these runs, the calibration set only contained one distinct response value, and thus predicted the same response value regardless of the predictor variable values.

Similarly with ordinal regression the values of κ_w decrease and the range of κ_w increases as the set size decreases. The analysis in Figure 4 shows that the current model and ordinal regression have similar performance with large calibration sets (128 and 64 points), but the current model outperforms ordinal regression with smaller calibration sets.

The wide range of κ_w for small sets appears to be influenced by the range of response values in the calibration set. There is moderate positive correlation (R = 0.54 and R = 0.78 for sets of eight and four, respectively) between κ_w and number of distinct response values in the calibration set. Figure 5 shows the results of the comparison of the model's performance for 'high' and 'low' calibration sets (see section 2.3.6) for runs of eight and four points[2].

[2] Note that for every run with 16 or more points, there were at least three distinct response values present in the calibration sets.

Even with only four points in the calibration set, agreement is relatively high (mean κ_w = 0.72), provided most response values are represented in the calibration set.

4. Discussion

Inductive modelling techniques based on Bayes' theorem have been applied previously in GIS to predict species distribution (Aspinall, 1992; Gorman et al., 2007), classify risk (Stassopoulou et al., 1998; Grêt-Regamey and Straub, 2006), model behaviour (Aalders, 2008) and to classify phenomena (Asadi and Hale, 2001; Corner et al., 2002). Our model similarly uses Bayes' theorem coupled with GIS to classify probability of crop and forage success. This research has compared the performance of Bayesian probability modelling with ordinal regression, and attempted to evaluate its performance with increasingly smaller calibration datasets.

Predictive models commonly require a large amount of data for calibration; the results reported above indicate that our model can provide useful results with very few trial data points, provided the trial data supply information for all or most response categories. However we recognise that in this case study the accurate results can partly be explained by the small number of variables used to develop the simulation, and therefore more case studies are needed using real world data.

Our case study included only three predictor variables, whereas an empirical application is likely to include more. In Section 2.3.2 we stated that predictor variables were chosen in consultation with forage experts at CIAT. Three additional predictor variables relating to soil were originally suggested (acidity, fertility and texture), however these were not included in the current case study because of the lack of accurate spatial layers for validation purposes.

The CaNaSTA method has been applied and evaluated in a limited number of real-world examples, as mentioned in the introduction. Atzmanstorfer et al. (2007), applying CaNaSTA to coffee and cowpea, found that the methodology provided insight into the interaction of agronomic and ecologic variables with environmental conditions, that was not previously available. They also found CaNaSTA to be an effective modelling tool in data sparse situations. Läderach et al. (2006) validated CaNaSTA for the case study of niche coffee growing regions in Colombia and Nicaragua. Comparing predicted high quality with evidence using the likelihood-ration chi square test, they found that CaNaSTA predicts niches likely to produce high quality coffee at $p = 0.014 – 0.081$ confidence levels.

Both Atzmanstorfer et al. (2007) and Läderach et al. (2006) used climatic, pedalogic and topographic attributes as predictor variables, including annual average precipitation, annual average temperature, dry months per year, annual average diurnal temperature range, mean annual solar radiation, dewpoint, soil type, elevation, aspect and slope. In order to be

usefully included in the model, spatial surfaces must be available for all predictor variables at the same resolution.

Care needs to be taken that conditional independence is not violated, and analysis can be undertaken to ensure that the most appropriate variables are included in the model. Future empirical research should examine the effects of inclusion or exclusion of predictor variables in the model.

The equations described here have been implemented with the framework of the SDSS CaNaSTA, however the discussion has been limited to the model itself, rather than the implementation of the SDSS.

5. Conclusion

CaNaSTA has potential as a tool for predicting ordinal responses of crops or other spatial phenomena, especially where calibration data is sparse. The results reported above are promising; however further testing is required on different simulated scenarios, and on real world scenarios. The analysis presented here was perhaps limited by using only three predictor variables. The inclusion of variables characterising the probabilistic nature of temporal events such as rainfall could improve both the model and its handling of temporal uncertainty. In applications of CaNaSTA to real-world modelling scenarios care needs to be taken in choosing the most appropriate predictor variables, depending on both the temporal and spatial scale of the response variable. CaNaSTA is particularly valuable for examining potential spatial response, and the reasons thereof, of specialised niche crops.

In this study, the model has been applied to a simulated dataset in order to allow controlled analysis of the impact of the size of the calibration dataset on the robustness of the model. Large validation sets, such as those used here, are not generally available in crop trial data. The purpose has been to test the model under controlled circumstances. The next step is to apply the model to real-world scenarios such as those mentioned above.

Future research and development aims to provide further analysis options for specialised crop response research, as well as the application of CaNaSTA to diverse spatial research problems.

Acknowledgements

The research described in this paper was funded in part by the Bundesministerium für Wirtschaftichle Zusammenarbeit und Entwicklung (German Federal Ministry for Economic Cooperation and Development, BMZ). The authors also acknowledge the support from the Deutsche Gesellschaft für Technische Zusammenarbeit (German Agency for Technical Cooperation, GTZ), the Diversification Agriculture Project Alliance (DAPA), the International Center for Tropical Agriculture (CIAT) and USAID. The majority of the research described here was undertaken at Curtin University of Technology.

References

Aalders, I. (2008). Modelling land use decision behaviour with Bayesian Belief Networks. *Ecology and Society*, 13(1), 16-37.

Asadi, H.H. and Hale, M.M. (2001). A predictive GIS model for mapping potential gold and base metal mineralization in Takab area, Iran. *Computers and Geosciences*, 27(8), 901-912.

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

Aspinall, R. and Veitch, N. (1993). Habitat mapping from satellite imagery and wildlife survey data using a Bayesian modelling procedure in a GIS. *Photogrammetric Engineering and Remote Sensing*, 59(4), 537-543.

Atzmanstorfer, K., Oberthür, T., Läderach, P., O'Brien, R., Collet, L. and Quiñonez, G. (2007). Probability modelling to reduce decision uncertainty in environmental niche identification and driving factor analysis: CaNaSTA case studies. In: Zeil, P. and Kienberger, S. (Eds.) *GeoInformation for Development – Bridging the Divide through Partneships*. Wichmann-Verlag, Heidelberg, pp. 33-43.

Barco, F., Franco, M.A., Franco, L.H., Hincapie, B., Lascano, C., Ramirez, G. and Peters, M. (2002). *Forrajes tropicales: base de datos de recursos genéticos multipropósito, versión 1.0 [CD-ROM]*, Centro Internacional de Agricultura Tropical CIAT, Cali, Colombia (Serie CD-ROM Publicación CIAT No. 329).

Bonan, G. (2002). *Ecological Climatology*. Cambridge University Press, Cambridge.

Bzreziecki, B., Kienast, F. and Wildi, O. (1993). A simulated map of the potential natural forest vegetation of Switzerland, *Journal of Vegetation Science*, 4, 499-508.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37-46.

Corner, R., Hickey, R.J. and Cook, S.E. (2002). Knowledge based soil attribute mapping in GIS: the Expector method. *Transactions in GIS*, 6, 383-402.

Fischer, H.S. (1990) Simulating the distribution of plant communities in an alpine landscape, *Coenoses*, 5, 37-43.

Fleiss, J. (1981). *Statistical Methods for Rates and Proportions*. Wiley, New York.

Gorman, J., Pearson, D. and Whitehead, P. (2007). Assisting Australian indigenous resource management and sustainable utilization of species through the use of GIS and environmental modelling techniques. *Journal of Environmental Management*, 86(1), 104-113.

Grêt-Regamey, A. and Straub, D. (2006). Spatially explicit avalanche risk assessment linking Bayesian networks to a GIS. *Natural Hazards and Earth Systems Science*, 6, 911-926.

Guisan, A. and Harrell, F.E. (2000). Ordinal response regression models in ecology. *Journal of Vegetation Science*, 11, 617-626.

Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G. and Jarvis, A. (2005). Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, 25, 1965-1978.

Jensen, F.V. (1996). *An Introduction to Bayesian Networks*. UCL Press, London.

Läderach, P., Vaast, P., Oberthür, T., O'Brien, R., Lara Estrada, L.D. and Nelson, A. (2006). Geographical analyses to explore interactions between inherent coffee quality and production environment. Paper presented at the 21st International Conference on Coffee Science, Montpellier, France, Sep 11-15.

La Morgia, V., Bona, F. and Badino, G. (2008). Bayesian modelling procedures for the evaluation of changes in wildlife habitat suitability: a case study of roe deer in the Italian Alps. *Journal of Applied Ecology*, 45, 863-872.

McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models, 2nd Ed*. Chapman & Hall, London.

Pearl, J. (1990). Bayesian decision methods. In: Shafer, G. and Pearl, J. (Eds.) *Readings in Uncertainty Reasoning*. Morgan Kaufman, San Mateo, pp. 345-352.

Press, W.H., Flannery, B.P., Teukolsky, S.A. and Vetterling, W.T. (1986). *Numerical Recipes: the Art of Scientific Computing*. Cambridge University Press, Cambridge.

Skidmore, A.K. (1989). An expert system classifies eucalypt forest types using Landsat Thematic Mapper data and a digital terrain model. *Photogrammetric Engineering and Remote Sensing*, 55, 1449-1464.

Spiegelhalter, D., Thomas, A., Best, N. and Lunn, D. (2005). *WinBUGS User Manual Version 2.10*. MRC Biostatistics Unit, Cambridge.

Stassopoulou, A., Petrou, M. and Kittler, J. (1998). Application of a Bayesian network in a GIS based decision making system. *International Journal of Geographic Information Science*, 12(1), 23-46.

Zhou, B., Zhang, X. and Wang, R. (2004). Automated soil resources mapping based on decision tree and Bayesian predictive modelling. *Journal of Zhejiang University Science*, 5:7, 782-795.

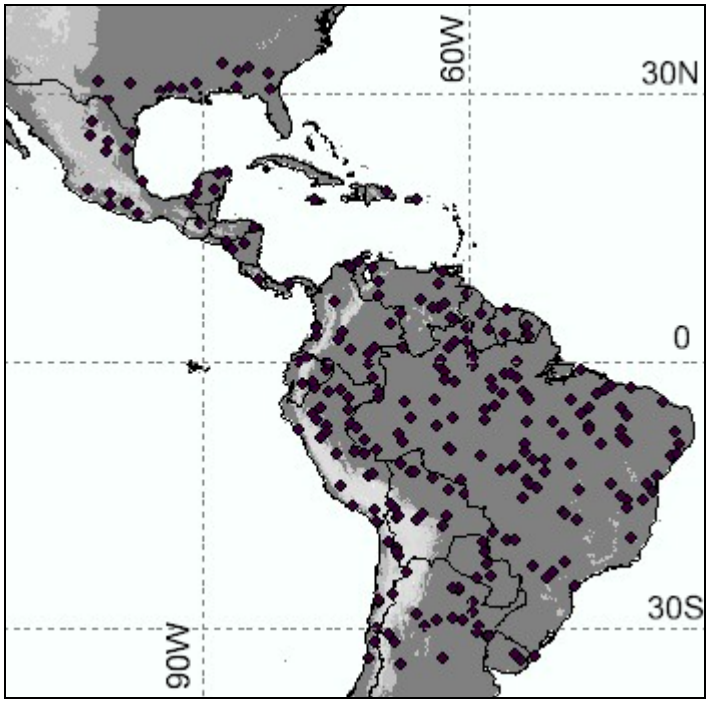

Figure 1. Location of simulated trial sites

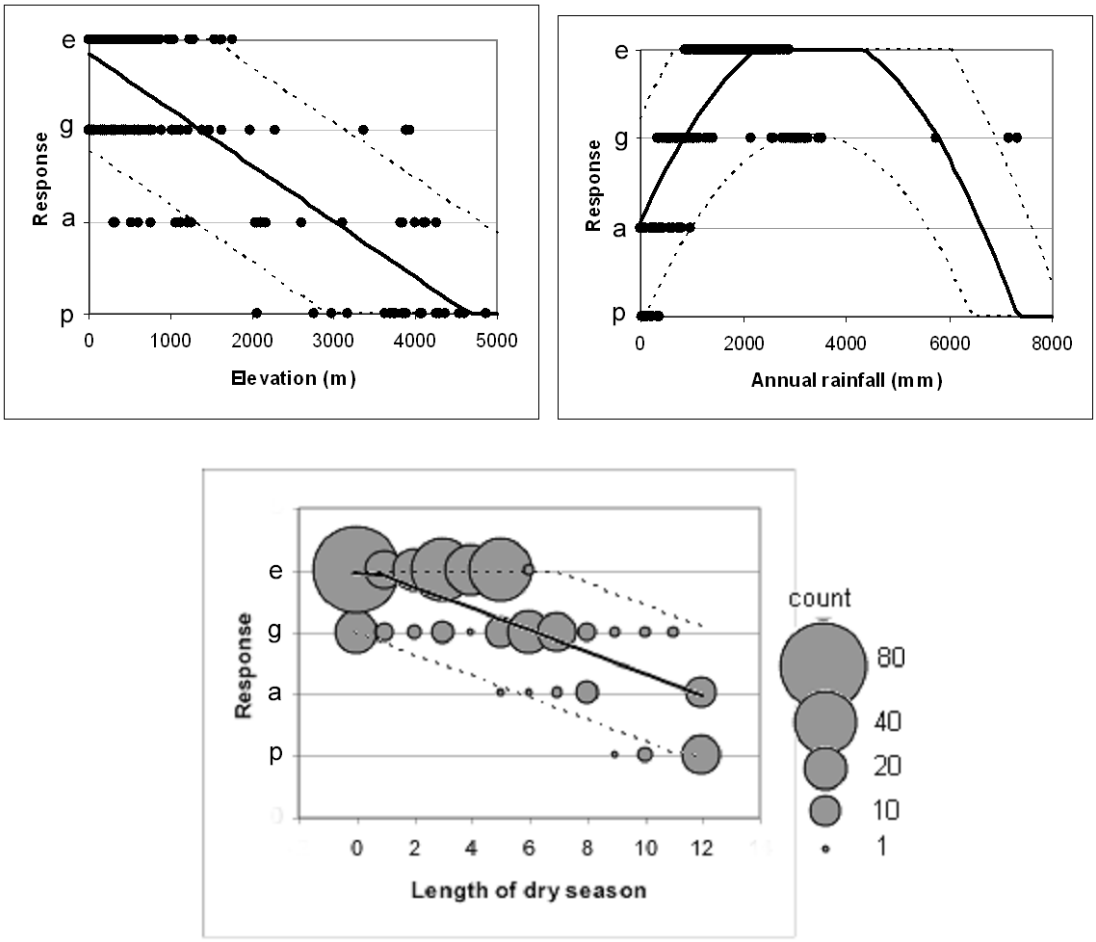

Figure 2. Response variables plotted against predictor variables (e = excellent, g = good, a = average, p = poor). Dots show data values; for length of dry season, size of bubbles shows number of records for each value. Solid line shows regression fit, dotted lines are 95% credible intervals.

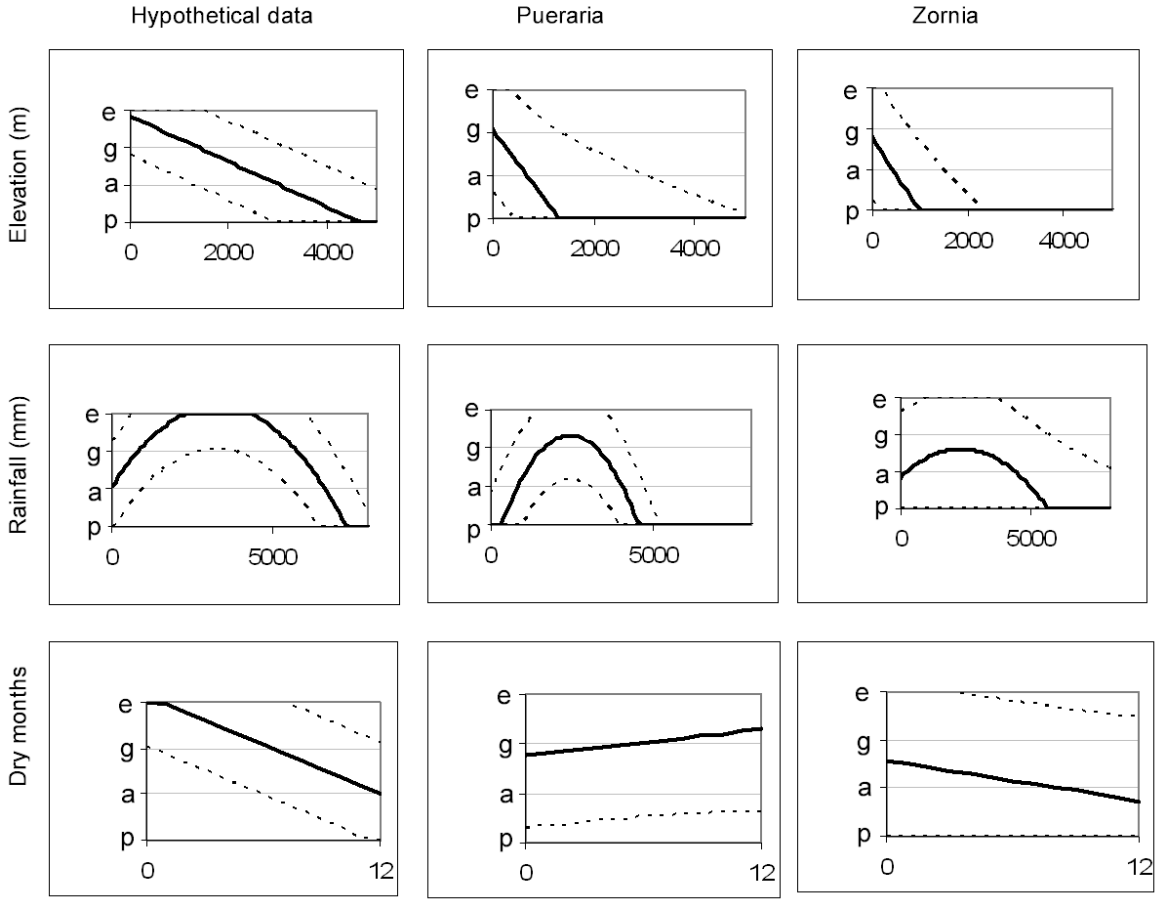

Figure 3. Regression fit (solid line) for simulated data (left), *Pueraria* (middle) and *Zornia* (right) for elevation (linear; top), rainfall (second order polynomial; middle) and dry months (linear; bottom) against response. Dotted lines are 95% credible intervals. The database has no trials in locations with elevation over 1370m or dry season longer than six months. Adaptation for all species is expected to be poor above these limits, reinforcing the negative trend for elevation and length of dry season.

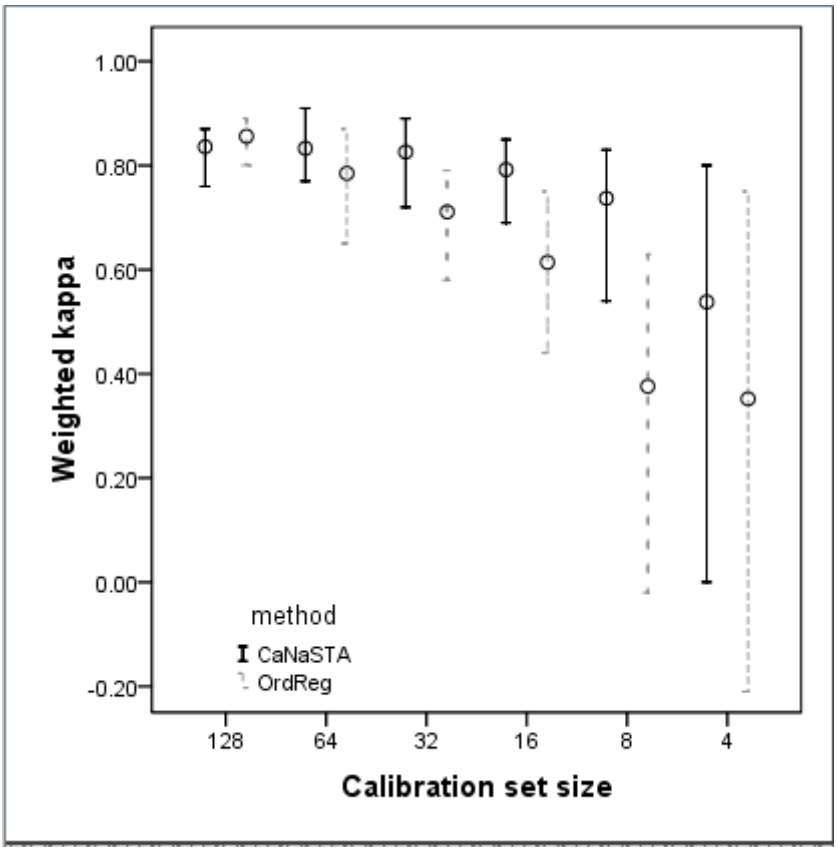

Figure 4. Summary of κ_w for ten runs on each set size for CaNaSTA and ordinal regression. Bars show maximum, minimum and average kappa values.

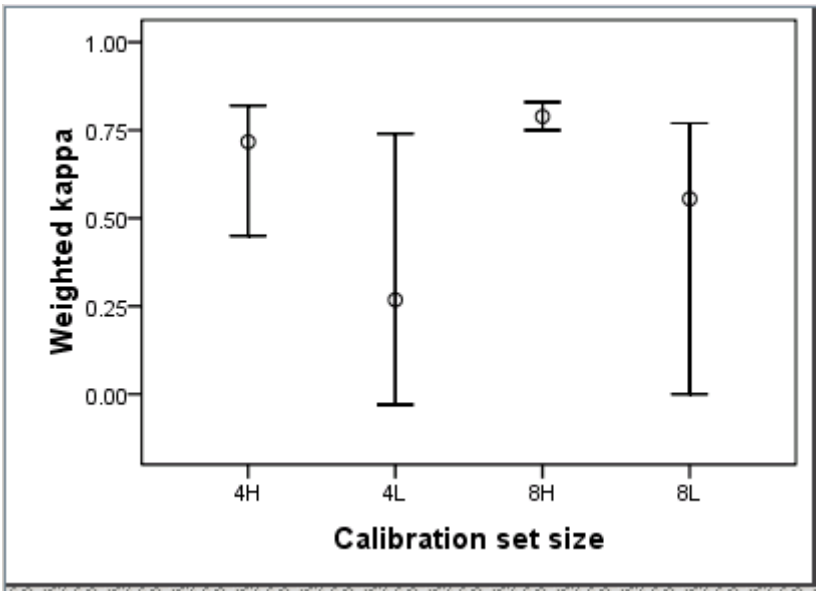

Figure 5. Summary of κ_w for runs of eight and four points. H/L = high/low proportion of response values represented in the calibration set.

Table 1. Correlation and joint information uncertainty (JIU) between predictor variables in the study area

Variables	Correlation	JIU
Elevation and Rainfall	-0.42	0.03
Elevation and Dry months	0.51	0.14
Rainfall and Dry months	-0.77	0.14

Table 2. Example trial data (classified values in brackets)

Site	Elevation (m)		Annual Rainfall (mm)		Length of Dry Season (months)		Adaptation	
1	28	(1)	1419	(4)	0	(1)	Excellent	(e)
2	549	(2)	546	(2)	9	(5)	Good	(g)
3	575	(2)	1850	(5)	5	(3)	Excellent	(e)
4	188	(1)	1715	(3)	5	(3)	Excellent	(e)
5	1470	(3)	903	(2)	7	(4)	Good	(g)
6	1207	(3)	265	(1)	12	(5)	Average	(a)
7	3839	(5)	364	(1)	9	(5)	Poor	(p)
8	763	(1)	1400	(4)	5	(3)	Excellent	(e)
9	4	(1)	1842	(5)	0	(1)	Excellent	(e)
10	28	(1)	1282	(4)	5	(3)	Excellent	(e)
11	12	(1)	1196	(3)	0	(1)	Excellent	(e)
12	4117	(5)	724	(2)	8	(4)	Average	(a)

Table 3. Conditional probabilities derived from Table 2. Raw frequencies (adjusted values).

Elevation class	**1**	**2**	**3**	**4**	**5**
$P(y_e \mid elev_i)$	1.00 (0.97)	0.50 (0.49)	0.00 (0.01)	0.00 (0.583)	0.00 (0.01)
$P(y_g \mid elev_i)$	0.00 (0.01)	0.50 (0.49)	0.50 (0.49)	0.00 (0.167)	0.00 (0.01)
$P(y_a \mid elev_i)$	0.00 (0.01)	0.00 (0.01)	0.50 (0.49)	0.00 (0.167)	0.50 (0.49)
$P(y_p \mid elev_i)$	0.00 (0.01)	0.00 (0.01)	0.00 (0.01)	0.00 (0.083)	0.50 (0.49)
Rainfall class	**1**	**2**	**3**	**4**	**5**
$P(y_e \mid rain_i)$	0.00 (0.01)	0.00 (0.01)	1.00 (0.97)	1.00 (0.97)	1.00 (0.97)
$P(y_g \mid rain_i)$	0.00 (0.01)	0.667 (0.653)	0.00 (0.01)	0.00 (0.01)	0.00 (0.01)
$P(y_a \mid rain_i)$	0.50 (0.49)	0.333 (0.326)	0.00 (0.01)	0.00 (0.01)	0.00 (0.01)
$P(y_p \mid rain_i)$	0.50 (0.49)	0.00 (0.01)	0.00 (0.01)	0.00 (0.01)	0.00 (0.01)
Dry season class	**1**	**2**	**3**	**4**	**5**
$P(y_e \mid dry_i)$	1.00 (0.97)	0.00 (0.583)	1.00 (0.97)	0.00 (0.01)	0.00 (0.01)
$P(y_g \mid dry_i)$	0.00 (0.01)	0.00 (0.167)	0.00 (0.01)	0.50 (0.49)	0.333 (0.326)
$P(y_a \mid dry_i)$	0.00 (0.01)	0.00 (0.167)	0.00 (0.01)	0.50 (0.49)	0.333 (0.326)
$P(y_p \mid dry_i)$	0.00 (0.01)	0.00 (0.083)	0.00 (0.01)	0.00 (0.01)	0.333 (0.326)