A BAYESIAN MODELING APPROACH TO SITE SUITABILITY UNDER CONDITIONS OF UNCERTAINTY

Rachel O’Brien1, 2, Simon Cook and Michael Peters

1International Center for Tropical Agriculture
Cali, Colombia

Robert Corner

2Department of Spatial Sciences
Curtin University of Technology
Perth, WA, Australia

ABSTRACT

Precision agriculture in the developed world tends to rely on the collection and analysis of significant amounts of data using sophisticated equipment. For smallholder farmers in the developing world this data is simply not available. However, data does exist that can be used to perform analysis at sub-farm level. This data includes spatial information in the form of geographical information databases on climate and soils, crop trial data, expert knowledge and farmer knowledge.

Often this data and knowledge is sparse and only reliable to a certain extent. Most statistical techniques are unworkable in this situation, with too many assumptions needing to be made in order to perform any analysis. Bayesian modeling techniques provide a simple yet robust way of combining existing data and knowledge in the form of probabilities, whilst keeping uncertainties in the data and knowledge explicit.

A tool called CaNaSTA (Crop Niche Selection in Tropical Agriculture) has been developed demonstrating these techniques for recommending forage crops to farmers and their advisors in Central America. This paper discusses the development of this tool and compares Bayesian modeling with conventional methods, and shows that Bayesian techniques provide the best model in the context of selecting suitable crops using sparse and uncertain data.

Keywords: Forages, decision support, uncertainty, risk, expert knowledge, developing countries, adoption, Precision Agriculture, GIS.
INTRODUCTION

Agricultural development plays a significant role in the economies of developing countries. Natsios (2001) stresses the importance of agricultural development as a means to alleviating poverty in developing countries. A relatively large proportion of the population in these countries is smallholder farmers. In developing countries, farmers are often poor and farm very small areas using rudimentary techniques. Solutions that have the potential to improve farmers’ situations are likely to be complex, consisting of a mix of technology, improved crop varieties, technical assistance, financial support and education.

In order to alleviate poverty and improve food and income security in the developing world, sustainable production systems are necessary, balancing environmental protection with social and economic sustainability. Intensification of production may be the only solution for resource-poor farmers (Peters et al., 2001). The production environment for smallholder farmers in the tropics is characterized by uncertainty and risk. It is clear that farming systems are complex, and decisions are influenced by biophysical, social, economic and market factors. Furthermore, any decisions made will, in turn, impact on biophysical, social, economic and market factors.

The role of forages in agricultural development was discussed in O’Brien et al. (2002). Despite the potential of improved forages, adoption has been slow, of legumes in particular. One barrier to adoption is lack of information on which forage species are suited to a farmer’s unique environment and for what reasons (Schultze-Kraft and Peters, 1997). As with other tropical crops, many uncertainties exist in these environments, and decision-making is often risky.

RISK AND UNCERTAINTY IN TROPICAL AGRICULTURE

Risk and uncertainty are factors in most decision-making, especially when the decision-making process has spatial aspects. In the case of supporting farmers’ decisions about forage selection, there are a number of sources of risk and uncertainty. For example, there is uncertainty for the farmer about which forage species exist, and what their properties are. Also, uncertainty exists about forage species’ management and environmental requirements. There will also be uncertainty due to climate variability and soil variability. These and other uncertainties expose smallholder farmers to risk (see O’Brien et al., 2002). Approaches to reducing uncertainty depend on its source and its type. In order to incorporate knowledge, uncertainty and risk in the modeling process, appropriate forms of modeling must be selected.

The problem of deciding what to plant where is inherently spatial, since crops and forages are produced within a spatially variable environment. The environment displays heterogeneity at different scales, for example, soils tend to vary at a finer scale than climate. Farmers are interested in species that will thrive in their particular location. Hence much of the uncertainty surrounding the selection of species is also spatial (and in fact spatio-temporal) in nature.

Decision-makers can be classed as risk-taking (sometimes termed risk-seeking), risk-neutral or risk-averse, depending on the level of risk they are
willing, or able, to accept. Attitudes to risk affect adoption of new crops, technology and practices.

**MODELING THE FORAGE SELECTION DECISION**

The purpose of modeling in this case is to predict the success of forage species in specific niches, with particular biophysical, socio-economic and management characteristics. Functional models support the tactical decision of the farmer regarding which forage species to select.

A model is only useful to the degree that it represents reality in a way that supports decision-making. Therefore the model needs to be consistent and accurate and, at the same time, reduce the complexities of the decision-making process to make them manageable.

Simply developing a model of species’ suitability is only part of the process. The results of the model need to be made available to the decision-maker. An appropriate way of delivering this information is by means of a Spatial Decision Support System (SDSS).

**Criteria for model selection**

The question of which species are suitable where can be phrased in a number of ways. The problem can be interpreted from an ecology perspective (natural occurrence of species) or from a niche modeling perspective (suitable cultivation of species). Essentially, all reduce to the same decision problem: “What is the likelihood that species $\alpha$ is suitable at location $\beta$?”

Although the purpose of ecological modeling is different from that of niche modeling, it is clear that there is significant overlap in their approaches. There is an extensive body of literature on ecological modeling that examines the question of where a plant or animal species might be found, based on environmental factors. The resulting models are known as habitat models or distribution models, and they relate the geographical distribution of species to their environment (Guisan and Zimmermann, 2000). Clearly, the decision of what to plant where is related to habitat modeling.

In selecting a model to address the problem of which forages to plant where in developing agriculture, a number of criteria should be considered. The first criteria are the ability to work with small datasets and expert knowledge and the ability to predict a range of species’ responses, rather than just ‘presence’ and ‘absence’. In addition, the model must display low structural complexity and must be easy to communicate and to implement spatially.

**Modeling approaches**

The main models considered in this research were logistic regression, Generalized Linear Models (GLM), Generalized Additive Models (GAM), Artificial Neural Networks (ANN), Classification and Regression Trees (CART), environmental envelopes, fuzzy rule-based methods and Bayesian probability models. Some of these methods are empirical, data-driven methods
and others are causal, knowledge-driven methods. Some techniques can be either, or a combination of both.

The criteria of working with small datasets and expert knowledge simultaneously dictate that models that can be at least partly knowledge-based are preferable. Models depending mostly on expert knowledge will generally display low structural complexity and ease of communication.

Logistic regression is a popular technique in habitat modeling and is capable of producing good results, providing a number of assumptions are met. Logistic regression is useful for predicting a binary response from either continuous or categorical predictors. Although logistic regression is a robust method, it is dependent on relatively large datasets. In selecting suitable species for farmers in the tropics, a method is needed that works with relatively sparse datasets.

GLMs and GAMs were first developed in the 1960s and have since been used extensively in ecological research (Guisan et al., 2002). GLMs are extensions of linear models, allowing for non-linearity and non-constant variance structures in data. GAMs are a further extension of GLMs, where the only underlying assumption is that the functions are additive and the components are smooth (Guisan et al., 2002). These models are particularly useful in ecology modelling because underlying data is usually highly non-linear and may take on many different distribution forms. In essence, GLM and GAM are extensions of logistic regression and therefore face many of the same issues regarding applicability to forage selection. The methods are well suited to ecological presence/absence data, in particular where little is known about the relationship between predictor variables and species’ presence.

ANN is an artificial intelligence technique based on a representation of the neural interactions in the human brain. Information is passed through a number of nodes, resulting in values or classifications. In tropical forage selection, sparseness of data and the requirement of being able to incorporate expert knowledge make ANN a less promising method. For crops or forages for which little is known about their ecological processes, but where a large dataset is available, ANN could be useful, but this situation is unlikely to realistically occur.

CART is derived from the concept of decision trees, with a decision taken at each node in the tree depending on the observation value, and with leaves of the tree representing resulting classifications. The amount of data required to specify robust trees is a drawback in the case of predicting species success for smallholder farmers in the tropics. However, the fact that expert opinion can be incorporated relatively easily is beneficial. In addition, trees can be easily interpreted for biological meaning. It is not clear, however, how to deal effectively with uncertainty in decision trees. CART could be a useful tool for organising data and incorporating expert knowledge.

Environmental envelopes, or habitat envelopes, define an envelope in multi-dimensional attribute space within which the species is expected to be found. A number of different algorithms have been used to define these envelopes. With environmental envelopes, the response variable tends to be presence/absence, with responses classified as ‘present’ (or ‘suitable’) if they fall within a given percentile for all variables. An additional class of
‘marginal’ is often added for responses that fall within that percentile for some factors but outside for others. The most basic environmental envelopes are rectilinear. However, more sophisticated algorithms have been developed for defining environmental envelopes, including minimum bounding polygons and fuzzy clouds. Environmental envelopes have already been used extensively for crop selection in tropical agriculture. Most databases on crop adaptation store environmental data in formats that lend themselves to mapping using envelopes, even if this approach was not initially intended. For example, EcoCrop (FAO, 2000) lists minimum and maximum environmental criteria for various crops, sometimes including marginal boundaries. Therefore, environmental envelopes are a highly suitable method for use in tropical forage selection.

Fuzzy logic (Zadeh, 1965) seeks to relax the crisp and deterministic classifications imposed by Boolean logic. Fuzzy membership generalises Boolean logic by assigning the value 1 to the state ‘true’, 0 to the state ‘false’ and allowing values between these two numbers. Fuzzy theory could be a valuable tool for forage selection. The many uncertainties, both in geographical and attribute space, could be addressed using fuzzy classification. Fuzzy logic provides a many-valued alternative to the binary nature of traditional Boolean logic, where values are true/false or presence/absence. This allows for classifications of ‘marginally suitable’, in addition to classifications of ‘not suitable’ and ‘suitable’.

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 simple Bayesian model defines prior and conditional probability distributions for each node and then uses combination rules to propagate conditional probability distributions through the network. The probability distributions may be derived from data, set by experts or defined from a combination of data and expert opinion. This process of combining probabilities produces conditional probabilities for each possible outcome. Where a node has multiple parents, a conditional probability table (CPT) is defined, defining probability distributions for each possible combination of parent node states. The decision problem is characterised by uncertainty, and probability-based methods are well equipped to incorporate uncertainty. In addition, the ability to incorporate both data and expert knowledge makes probability-based modelling appealing. Bayesian models can have varying degrees of complexity, but even complex models have clear biological meaning. Probability-based modelling is a promising approach to the problem of selecting tropical forages.

In summary, for the decision problem of selecting forages, environmental envelopes are promising, and have already been spatially implemented in many habitat distribution problems. Fuzzy rule-based methods deal well with uncertainty and expert knowledge. Finally, Bayesian probability methods allow the combination of both data and knowledge and handle uncertainty well. It is proposed that a simple Bayesian probability model is well suited to the decision problem. It is envisaged that environmental envelope concepts
and fuzzy rule-based methods display some overlap with a spatial implementation of a simple Bayesian model.

**Probabilistic GIS modelling**

Let $X$ be an independent (predictor) variable and $Y$ be a dependent (response) variable, and let $y_i$ be a possible state of $Y$ and $x_j$ be a possible state of $X$. The conditional probability $P(Y = y_i \mid X = x_j)$ denotes the probability of $y_i$ being the state of $Y$ given that $x_j$ is the state of $X$. For simplicity this can be written as $P(y_i \mid x_j)$. Then the posterior probability of the response variable being in a certain state, given that the predictor variables are in certain states, can be calculated using the following equation:

$$P(y_i \mid x_{j1}, x_{j2}, \ldots, x_{jk}) \propto P(y_i) \prod_{k=1}^{l} \left( \frac{P(y_i \mid x_{jk})}{P(y_i)} \right)$$

Given all possible combinations of $P(y_i \mid x_{jk})$, for all $i$, $j$, $k$, a full conditional probability table (CPT) can be created (Figure 1). The proposed method is essentially to combine these predictor variables to define probabilistic environmental envelopes based on Bayesian modelling techniques. The Bayesian model employed will be the most simple Bayesian network possible, with all input variables at one level, feeding simultaneously into the output variable (Figure 2). The inputs are the predictor variables, including climate and soil characteristics. The output is the probability of forage suitability under these conditions, measured for example as probability of adaptation.

![Figure 1. Populating a CPT for 6 conditionally independent variables $X^k$ with 5 states each and a dependent variable $Y$ with 4 states.](image-url)
It is not necessarily the case that all variables act simultaneously – for example, in reality, elevation may affect both temperature and soil characteristics, and temperature itself may affect soil characteristics, so that a more complicated network could be more valid. However, sparse datasets make defining such a structure more problematic. In addition, a simple structure makes both implementation and explanation of the model much more straightforward.

In the case of selecting forage species for a specific niche, the problem is not just to predict the success of one species, but of many species simultaneously at one location, in order to identify the most promising species. Therefore, the model needs to apply some sort of ranking or filter to identify a subset of species for consideration at any one location. Filters can be applied by removing from consideration any species which do not meet tolerance or use requirements, if these are defined. Ranking can be applied once posterior probabilities have been derived for each variable for the conditions at the location under consideration.

From posterior probabilities, a ranked ‘basket of options’ can be selected for the location in question. At the same time, uncertainty information is retained, based on the database, expert knowledge and probability combinations. Sensitivity analysis can also be performed by inspecting how much probability distributions change, depending on changes in variable states. Posterior probability distributions, uncertainty information and sensitivity information can all be communicated using maps and graphs.

The model allows information from diverse sources on success of forages to be combined to predict success distributions for any combination of variables. The model incorporates uncertainty, retaining uncertainty information throughout the model and allowing this information to be displayed and interrogated in a GIS environment. The next step is to formalize this model as a spatial decision support system.
SPATIAL DECISION SUPPORT SYSTEM

The purpose of a DSS is to provide data, procedures and analytical capability leading to better-informed decisions. Typically, a DSS consists, therefore, of data, a rule-base, algorithms for combining these and a user interface. A DSS is not necessarily computer-based, but in this discussion it is assumed a DSS is implemented in a personal computer environment. A spatially enabled DSS requires spatial input data, spatial analysis capabilities and/or spatial output. Because a considerable amount of uncertainty stems from spatial variation, a spatial DSS (SDSS) can provide the information necessary to manage some of this uncertainty. The spatial component is usually implemented using GIS technology.

A spatial forage DSS would combine data, knowledge and appropriate algorithms to provide recommendations to farmers (Figure 3). However, rather than just providing recommendations, it is important that the SDSS support decisions appropriately by providing as much relevant information as possible. This information can include maps and graphs of species’ suitability and species factsheets. In addition, uncertainties can be explicitly shown in both the maps and the graphs.

The remainder of this paper describes the development and implementation of a spatial forage DSS.

Figure 3. A spatial forage decision support system.
Overcoming potential barriers to uptake

A number of barriers exist to the successful uptake of DSS in agriculture, particularly in the developing world (Cox, 1996; Walker, 2002). Those relating to design considerations include complex design and presentation of DSS, unrealistic requirements for monitoring data, the irrelevance and inflexibility of many DSS, lack of user confidence and poor data availability and quality.

The problem of poor data availability and quality is overcome partly by allowing uncertainty to be explicit in the structure of the model itself. Also, the incorporation of expert knowledge supplements unavailable data. There are no requirements for the farmer to provide monitoring data as input to the DSS. The SDSS is implemented as a stand-alone piece of software, meaning there is no requirement for other software to be present on the user’s computer (such as proprietary GIS or database software).

The design and presentation of the SDSS is intended to be simple and functional. At the most basic level, users need only be able to interpret maps in order to select their location. Displaying information in a number of ways (maps, graphs and numerically) should assist the user with understanding the outputs of the model in order to make better-informed decisions. Lack of user confidence can be addressed by reducing and describing all sources of uncertainty in the model and by allowing the user’s own knowledge to be incorporated.

This implementation therefore allows for sparse and uncertain data, works with expert knowledge and deals with uncertainty. In addition, the concept of the modeling process is fairly intuitive to follow, without the need for the user to understand the equations. Therefore, a user can decide how much faith to put in the information, because the entire process is transparent. This implementation attempts to present accurate results, but at the same time concentrates on delivering a structurally uncomplicated model and providing results that are straightforward in their interpretation. This process therefore addresses most of the problems encountered with other agricultural DSS and SDSS.

Implementation

Software has been developed based on the research discussed here. The software is called CaNaSTA (Crop Niche Selection in Tropical Agriculture) (canasta is Spanish for basket, and the tool aims to offer a basket of options to farmers, particularly in Spanish-speaking Central America). CaNaSTA recommends species for a given location and situation and recommends locations for a given species. In addition, users can update data interactively and examine results through maps, tables and graphs.

Incorporating spatial capabilities into an agricultural DSS, as in CaNaSTA, facilitates data input, allows more informative output of results and allows spatial variability to be made explicit, both of results and of uncertainties related to the results.
An overview of the SDSS is shown in Figure 4. For further details on CaNaSTA, see the poster presented at this conference titled “CaNaSTA: a Spatial Decision Support System for Crop Niche Selection in Tropical Agriculture”

**Discussion**

The model can be assessed in a number of ways. The first is to check the accuracy of the functional model by comparing results from the SDSS with results from other sources. A second assessment is the comparison of the process of decision-making using the SDSS with other methods of addressing the decision problem. Finally, an assessment is needed of how well the SDSS meets the stated objectives.

Accuracy of the model was checked by comparing results from CaNaSTA with results from a number of other sources. CaNaSTA showed reasonable agreement with these sources when recommending species for a given location. For five selected species, CaNaSTA showed moderate agreement with experts regarding suitability in selected locations. Spatial comparisons were also made for selected species by visually and analytically inspecting maps produced from different sources, and CaNaSTA showed moderate agreement spatially with other sources.

CaNaSTA provides some benefits over existing systems for determining which forage species are suitable where. CaNaSTA provides adaptation distributions, ranks suitable species and produces dynamic maps of species’ suitability.

The development of the probabilistic GIS model and its development as an SDSS meet the objective of providing decision support in risky and uncertain environments. The appropriate reduction and description of different types of uncertainty allow farmers to better manage the risks associated with decision-making in uncertain environments. This is achieved in the implementation of CaNaSTA.

![Figure 4. Overview of CaNaSTA](image_url)
Providing information to farmers to support their decision-making in uncertain environments is a meaningful goal. Farmers want more and better information and extension workers want to find ways to provide this information. Scientists often have this information and want to find meaningful and consistent ways of providing it to those who could benefit from it. This research has contributed to these goals.

CONCLUSIONS

The research shows that even with limited data and knowledge, results can be obtained that support the farmers’ decision-making process. When uncertainties are made explicit, farmers can then make less risky decisions by taking these uncertainties into account. Providing access to decision support through a Spatial Decision Support System, such as CaNaSTA, ensures that the information is delivered in a consistent and robust manner. Trial data and expert knowledge previously inaccessible to farmers are made available so that decisions taken are better informed.

These decisions will increase the adoption of appropriate forages, contributing towards sustainability, improving meat and milk quality, combating food problems and, ultimately, improving the livelihoods of smallholder farmers and their communities in the developing tropics.

REFERENCES


