Spatial Driving Factor Analysis for Specialty Crops

Rachel O’Brien

School of Environmental Sciences
Charles Sturt University, Albury, NSW, Australia
Phone: +61 2 6051 9641 Fax: +61 2 6051 9897
Email: robrien@csu.edu.au

Presented at SIRC 2006 – The 18th Annual Colloquium of the Spatial Information Research Centre
University of Otago, Dunedin, New Zealand
November 6th-7th 2006

ABSTRACT

A spatial decision support system (SDSS) has been developed that combines data from trial sites with expert knowledge and biophysical data to predict where different crops may be successful. The tool is proving useful for the analysis of highly specialised crops, such as specialty coffee, where little trial data is available and variables influencing crop response are crop-specific. The SDSS provides maps showing the probability of success and the certainty level of the prediction in each location. Researchers require not only maps showing where niche crops are likely to succeed, but also the reasons behind predicted success and failure. Driving factors are interpreted as the most important variables in a model, and analysing these can provide this information. Simple driving factor analysis has been included in the SDSS, but methods are needed to determine combinations of driving factors within a model, which may be spatially variable throughout a study area. This paper discusses the present status of the SDSS, and explores ways of developing and implementing site-specific driving factor analysis.

Keywords and phrases: driving factors, spatial analysis, SDSS, CaNaSTA, specialty crops, specialty coffee

1.0 INTRODUCTION

Farmers throughout the world need to know that they are planting the right crops in the right locations. Some crops are broadly adapted, and, with adequate management, can thrive in a wide variety of conditions. Other crops, however, are highly localised, and will only be successful under very specific conditions. Environmental conditions can vary greatly even over small areas. For example, small-scale coffee farmers in the mountains of southern Colombia can add unique qualities to their products by selecting precise environmental conditions, such as slope, aspect and soil type. This is an important contribution to their ability to develop products with exclusive qualities that will be capable of fetching premium market prices (CTA, 2006).

Various modelling approaches exist to identify suitable niches for specific crops, and one such approach has been used to create a Spatial Decision Support System (SDSS), called CaNaSTA (Crop Niche Selection for Tropical Agriculture) (O’Brien, 2006; O’Brien et al., 2005). This tool assists decision-makers in identifying which crops are most suitable in different locations.

However, researchers and farmers are not simply interested in mapping where something will grow, but also the question of why these locations are suitable, and not others. Extending the existing SDSS to analyse driving factors will help answer this question. Driving factors are the most important variables in a model, and can vary significantly in different locations. For example, in one location on a coffee farm, slope and aspect may the most important variables, but at lower altitudes drainage might be the dominant variable. Understanding these driving factors will allow farmers to select the most suitable crops for different locations, and also allow promising new locations to be found for high-value crops.

The CaNaSTA algorithms have been tested in a number of contexts. The application of CaNaSTA to improved forage species selection was validated by comparing the results to outputs from different models as well as expert knowledge (O’Brien, 2004). Initial validation with computer-generated data indicates that CaNaSTA can provide useful results with very few trial data points (< 10), provided the trial data supply information for all
response categories (unpublished data). Further case studies applying CaNaSTA to coffee and cowpea are discussed briefly later in this paper.

2.0 BACKGROUND

Models to predict the response of agricultural crops have long been employed in agricultural research. Agricultural Decision Support Systems (DSS) have been in existence since at least the mid 1970s. There are hundreds of DSS available, covering production decisions relating to crops such as cotton, wheat and pasture (McCown et al., 2002). Increasingly, agricultural DSS are being implemented spatially, and a number of spatial models also exist which can be applied to crop performance modelling (O’Brien, 2004).

In order to effectively determine spatially variable driving factors from a model, the model needs to base spatial predictions on real cause-effect relationships. Most predictive distribution models are empirical, statistical models (Guisan and Zimmermann, 2000). However, in this case the question is not only the spatial distribution, but also the factors driving this distribution in different locations.

CaNaSTA was initially developed as a tool to suggest niche forage species to smallholder farmers in the tropics. Recently, it has also been applied to coffee quality analysis, cowpea performance on tropical hillsides and carbon concentration in soils (Atzmanstorfer et al., 2006).

2.1 CaNaSTA Algorithms

The engine used to develop CaNaSTA is Bayesian probability modelling (see for example Pearl, 1990). A Bayesian model was chosen because these methods allow for the combination of trial data and expert knowledge as model inputs, and handle uncertainty well. Uncertainty is a concern particularly when a small number of data points are available to specify the model. CaNaSTA attempts to address this issue by keeping the structure of the model simple and explicitly stating the certainty values associated with the model outputs.

A simple Bayesian model defines prior and conditional probability distributions and combines these to calculate posterior probabilities for each possible outcome. Conditional probability can be calculated from prior and joint probability:

$$P(Y | X) = \frac{P(Y, X)}{P(X)}$$

where $X$ is a predictor variable, $Y$ is the response variable, $P(Y | X)$ is the conditional probability of the response given the predictor, $P(Y, X)$ is the joint probability and $P(X)$ is the prior probability.

It can be shown that posterior probability can be calculated from conditional and prior probabilities:

$$P(Y | X^1, X^2, ..., X^n) \propto P(Y) \prod_k \left( \frac{P(Y | X_k^1)}{P(Y)} \right)$$

where $X_k$ is the $k^{th}$ predictor variable ($k = 1, ..., n$).

CaNaSTA calculates the posterior probability distribution for each grid cell in the study area. The posterior probability distribution shows the probability that the response variable falls into each class. For example, if the response variable is quality, and there are three possible response classes (‘high’, ‘medium’ and ‘low’), then the probability distribution can be denoted as $[P(Y = \text{‘high’}), P(Y = \text{‘medium’}), P(Y = \text{‘low’})]$, with the probability values summing to 1. CaNaSTA then calculates a number of metrics, which can be displayed as maps over the area of interest. The score metric is a weighted average of the values in the posterior probability distribution, devised as a way of displaying the entire probability distribution in summary on one map. The main assumption is that the classes are ordinal, and class $j$ is ranked higher than class $j - 1$. The probability of each individual response class can be mapped, and also the most probable response class in each location. In addition, certainty and stability metrics are calculated for each location, and can also be mapped.

Each conditional probability distribution is assigned a certainty value of ‘low’, ‘medium’ or ‘high’, depending on the number of relevant trial sites or the certainty of the expert knowledge. When calculating posterior probability, these are assigned the values of 0, 1 and 2 respectively, and simply averaged over predictor variables.
to produce a combined certainty value. A better approach would be to assign continuous fractional values between 0 and 1 to denote certainty (0 = complete uncertainty, 1 = complete certainty), but this has not yet been implemented in CaNaSTA.

Stability is evaluated from changes in the response distribution as the states of predictor variables change. Some variables may be more sensitive than others. Stability is computed by calculating in how many instances the response distribution changes significantly if only one variable changes into an adjacent class. Firstly, the difference score is calculated when each variable changes to an adjacent class. This yields two values for each predictor variable (one in each direction). These are averaged to give a measure of stability for each predictor variable.

### 2.2 Driving Factors

Driving factor analysis attempts to identify the variable classes that disproportionately contribute to high values (positive driving factors) and low values (negative driving factors) in the probability surface. In CaNaSTA, driving factors are calculated once a probability map (the score map, or the probability map of an individual response class) has been created.

A sample of size \( n \) is taken from a region of interest and sorted by response value so that three sets can be obtained:

\[
N = \text{the set of all elements in the sample size (size } n) \\
Q_1 = \text{the set of elements in the upper quartile, ranked on response (size } n(Q_1) = n/4) \\
Q_4 = \text{the set of elements in the lower quartile, ranked on response (size } n(Q_4) = n/4)
\]

For each predictor variable, the following can be calculated:

\[
n(x_i) = \text{the number of elements in } N \text{ that are in class } i \text{ for predictor variable } x \\
n(x_i, Q_1) = \text{the number of elements in } Q_1 \text{ that are in class } i \text{ for predictor variable } x \\
n(x_i, Q_4) = \text{the number of elements in } Q_4 \text{ that are in class } i \text{ for predictor variable } x
\]

Then class \( i \) for predictor variable \( x \) is considered a positive driving factor if:

\[
\frac{n(x_i, Q_1)}{n(Q_1)} \times \frac{n(x_i)}{n} \geq c
\]

and is considered a negative driving factor if:

\[
\frac{n(x_i, Q_4)}{n(Q_4)} \times \frac{n(x_i)}{n} \geq c
\]

where \( c (> 1) \) is a user-defined threshold, with default value of 2.0. That is, if class \( i \) of predictor variable \( x \) is disproportionately represented in either the upper or lower quartile, then it is a driving factor. Although the default for the two subsets is upper quartile and lower quartile (i.e., 25%), this value can also be user-defined.

A sample is chosen simply because analysing every grid cell becomes costly for large grids. The samples are selected at regular intervals across the region of interest, and a large enough sample should be chosen to accurately represent the variation within the region of interest. Because the sample size is also user-defined, if runtime is unimportant then the above analysis can be carried out over the entire region of interest.

To illustrate the driving factors calculation, say there are \( n = 100 \) locations in a sample, of which \( n(x_i) = 20 \) are in predictor variable class \( i \), and there are \( n(Q_1) = 25 \) locations in the upper quartile, of which \( n(x_i, Q_1) = 15 \) are in predictor variable class \( i \), then the left-hand side of equation 3 evaluates to 3.75, and class \( i \) of the predictor variable is therefore a positive driving factor.

Figure 1 shows a screenshot of the current driving factor analysis in CaNaSTA. The user sets a number of options which determine the values of \( n, n(Q_1), n(Q_4) \) and \( c \) in equations 3 and 4 above. The SDSS then reports on the driving factors found. In Figure 1 it can be seen that the variable called VERA_SRTM_4 (an elevation grid) is a positive driving factor when elevation is between 803 and 1645m. From the graph it can be seen that once elevation rises above 1646m, it becomes a negative driving factor.
In an early attempt to analyse combinations of driving factors, an option has been included to keep a selected factor constant and repeat the analysis. In this case, the model will only examine locations in the specified variable class, and analyse all remaining variables to determine additional driving factors.

Continuing the example shown in Figure 1, keeping elevation between 803 and 1645m constant, the next most significant driving factor is annual rainfall between 1800 and 2422mm. Figure 2 maps the score value for those locations where elevation was kept constant as a driving factor, but does not map any new driving factor information.

This serves to illustrate the limitations of the driving factor analysis implemented in CaNaSTA to date, namely the inability to consider the effect of combinations of driving factors easily. However, it is a starting point for mapping distinct combinations of driving factors.

Figure 1. Screenshot of driving factor analysis in CaNaSTA
Figure 2. Map of score values for the entire region (top) and for only those locations with elevation between 803 and 1645 m (bottom). The bottom map only shows those locations where elevation is already known to be a driving factor, allowing further analysis to be carried out only in this region.
3.0 CANASTA APPLICATIONS

3.1. Specialty Coffee

The International Center for Tropical Agriculture (CIAT), is investigating alternatives for diversifying smallholder coffee agro-ecosystems (Atzmanstorfer et al., 2006), in particular specialty coffee. A case study was carried out linking coffee quality data from 88 farms to ten climatic, topographic and pedologic factors in the Cauca province in southern Colombia. CaNaSTA was used to identify ecological niches with high probabilities of producing superior quality specialty coffee (Figure 3).

In this case study, researchers found that no single factor drives coffee quality, but rather a combination of specific variables and specific ranges within these variables. For example, CaNaSTA found elevation between 1800-2000m and annual median temperature between 16-18C to be positive driving factors in the study area, whereas annual median temperatures between 19-20C proved a negative driving factor (Atzmanstorfer et al., 2006). However, the combined effects of driving factors were difficult to ascertain.

3.2. Cowpea

CIAT has also identified forage seed and leaf meals produced from cowpea (*Vigna unguiculata* (L.) *Walp.*) as a possible way for many smallholder farmers to substitute commodity coffee cultivation (Atzmanstorfer et al., 2006). CaNaSTA was used to identify optimal growing areas for cowpea grain and biomass production. Seven trial sites were established in the Cauca province of Colombia and four different cowpea accessions were trialled.
at these sites. Probability maps were then produced from these trial data using CaNaSTA and driving factors subsequently analysed. CaNaSTA identified a number of different driving factors for cowpea, including warm temperatures, low altitudes, a low diurnal temperature range, and a high dew point (Atzmanstorfer et al., 2006).

Again, it was difficult to ascertain the combined effects of these driving factors using CaNaSTA in its present state.

As mentioned in section 2.2, currently CaNaSTA has limited capacity to map driving factors. Certainly, specific ranges of a variable denoted as a driving factor can be mapped (see Figure 2). However, researchers require a more sophisticated approach, with different areas coded by their one or two most important driving factors. Therefore an approach is needed to not only determine driving factors, but also combinations of driving factors at each specific location.

4.0 APPROACHES TO DETERMINING DRIVING FACTORS

Although the approach outlined in section 2.2 above provides an indication of driving factors in a model, the results are global within the study area and not site-specific. Furthermore, although the impact of individual driving factors can be derived, it is not straightforward to determine the impact of combinations of factors. An initial strategy is to hold the most significant driving factor constant, and reapply the algorithm to discover the next most significant driving factor (see section 2.2).

Sensitivity analysis (see for example Dittmer and Jensen, 1997) provides methods for analysing which variables are the most sensitive in a model, in other words, changing the variable will have a large impact on the results of the model. CaNaSTA currently includes a stability metric based on sensitivity. If a variable is stable across all classes, then changing its class has little effect on the model, and the variable cannot be a driving factor. If, on the other hand, changing a variable’s class impacts the probability distribution of the response variable, then that variable has the potential to be a driving factor. Further research is needed to determine whether this approach will be useful in determining combinations of driving factors.

Moving windows approaches (see for example Wu et al., 2006) analyse a study area by focussing on a small region, and moving the focus (“window”) across the study area until the entire region has been examined. Moving windows could be used to analyse driving factors locally rather than globally across the study area. The size of the window applied is likely to influence results. If the window is too small, then it is possible the results will be overly fragmented and difficult to interpret. If the window is too large, then useful spatial detail could be lost. In addition, windows need to overlap to avoid unwarranted boundary effects when moving from one analysis window to the next. Again, further research is needed to verify this approach to determining site-specific driving factors.

Geographically weighted regression models (see for example Brunsdon et al., 1999) are designed to avoid the assumption that the same relationships in processes occur throughout the region. These methods could provide another approach to analysing driving factors locally rather than globally. There is some evidence that driving factors may be non-stationary in applications of CaNaSTA. For example, exploratory analysis on the coffee data shown in Figure 3 above suggests that the driving factors for specialty coffee quality are different in Inzá than in other regions.

The result of this driving factors analysis needs to be a map, and associated attributes, that is easily interpreted, showing which combinations of driving factors are most important in which locations.

5.0 CONCLUSIONS

Algorithms to accurately determine site-specific driving factors can help researchers in a number of ways. Maps will highlight which variables are important for the model in different locations. If a particular variable is consistently absent from the set of driving factors at all locations, then it can be concluded that that variable need not be included in the model. These maps will also quantify the level of homogeneity or heterogeneity of the study region. If combinations of driving factors tend to cluster in different locations, then it may be a valid conclusion that the study region should be split into subregions for further analysis.

General driving factor analysis has already been included in CaNaSTA and researchers are finding this information useful. However, it is clear that further research is needed on how best to determine site-specific combinations of driving factors and how to present and interpret this information.
ACKNOWLEDGEMENTS

The research described in this paper was funded in part by the Bundesministerium für Wirtschaftliche Zusammenarbeit und Entwicklung (German Federal Ministry for Economic Cooperation and Development, BMZ). The author also acknowledges 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.

REFERENCES


