I Encontro de Modelagem
ESTATÍSTICA

PRECISION AGRICULTURE USING STATISTICS

Ana Julia Righetto[1], Thiago G. Ramires[2] e Luiz R. Nakamura[3]

[1]Instituto Agronômico do Paraná, Londrina, Paraná, Brazil
[2]Departamento de Matemática, UTFPR, Apucarana, Paraná, Brazil
[3]Departamento de Informática e Estatística, UFSC, Florianópolis, Santa Catarina, Brazil

ABSTRACT

The cultivation of sugar cane has been gaining great focus in several countries due to its diversity of use. The modernization of agriculture has allowed high productivity, which is affected by the invasion of weeds. With sustainable agriculture, the use of herbicides has been increasingly avoided in society, requiring more effective weed control methods. In this paper, we used the multinomial logistic regression model that can be used as a classification tool, in the sense that it models a discrete response variable with more than two possible outcomes in a nominal scale. With these model it was possible to identify the invasion of weeds in the field, using four color spectra as regressor variables obtained by a multispectral camera mounted on an unmanned aerial vehicle. With the exact identification of the weed infestation, it is possible to carry out the management in the field with herbicide applications in the exact places, thus avoiding the increase of the cost of production or even dispensing with the use of herbicides, effecting the mechanical removal of them.

keywords: GAMLSS; multinomial logistic regression; modern agriculture; statistical modelling; weed management.

1 INTRODUCTION

Weed infestation is one of several problems involved in the cultivation of sugarcane. Weeds are plant species present in areas of human intervention, which are unwanted and often contrary to the goals of those who changed the primitive environment. They represent the energy imbalance created with the disturbance of the environment and can affect both the output and the quality of the products, since the weed competes for water and nutrients.

The introduction of informatics as a tool for the management and simulation of agricultural operations was one of the facts that had the greatest impact on the reduction of production costs of sugar cane in Brazil. Programs and systems developed for this purpose allowed the reduction in the fleet of trucks, tractors, harvesters and implements, maximization of the sugar quantity per hectare, optimization of the operation of the harvest fronts, evaluation of online performance and control of all operational agricultural activities. Adopting the method of soil mappingin regular grid allows the producers, who use the localized application of fertilizers,to make agribusiness more competitive and efficient in agricultural management and productivity increase. So the purpose of the work is to use the multinomial logistic regression model to differentiate what is soil, what is weed and what is actually sugar cane in a field of study.

2 METHODOLOGY

The data set used in this paper was collected in an experimental farm, situated in Piracicaba, São Paulo, Brazil, whichin the plantation was divided into two experimental fields, named field 1 and field 2. For both fields, the plantation was in the third ratoon, with 120 days after its last harvest. An unmanned aerial vehicle was used in order to collect the data, in which a multispectral camera was coupled taking pictures of both fields and returning the following information: level of near infrared band (NIR); level of red edge band (RE); level of red band (R); level of green band (G); latitude coordinate (lat); and longitude coordinate ($long$).

After obtaining the color bands and coordinate information, researchers went to the fields and registered manually, for some coordinates, what was the response variable Y, that is divided in three levels: `1`: soil, `2`: sugar cane or `3`: weed. It is noteworthy that level `3` groups different weed species: *Brachiaria decumbens*, *Cynodon dactylon* and *Amaranthus viridis*. The total points observed by the drone for both fields was $N = 127,853$, of which only $n = 8,801$ (6.88%) were manually classified, that is, only 8,801 observations of the data set contain information about the response Y.

The main idea here is to model Y based on the information of NIR, RE, R and G using a statistical model. Later, we use the fitted model to predict all the 119,052 unobserved values of Y, making it possible to better characterize weed invasion and providing a powerful tool to select in which places the application of herbicide or mechanical removal is needed. Further, the same model can be used to make predictions in new plantations, making that new data collection in the field is no longer required.

The multinomial logistic regression model can be used as a classification tool, generalizing the logistic regression, in the sense that it models a discrete response variable with more than two possible outcomes in a nominal scale, i.e. there is no specific ordering for the response. The goal here is to model the odds of the response as a function of a set of explanatory variables [1].

Basically, we may consider k outcome levels, in which one level is chosen as the referent or baseline level and the other $k - 1$ outcomes are separately regressed against the referent outcome. For $k = 3$, the multinomial logistic regression model is given by

$$P(Y = 1) = \frac{\mu}{1 + \mu + \sigma}, \; P(Y = 2) = \frac{\sigma}{1 + \mu + \sigma} \text{ and } P(Y = 3) = \frac{1}{1 + \mu + \sigma}. \tag{1}$$

where μ and σ represent the odds ratio between the levels $Y = 1$ versus $Y = 3$ and $Y = 2$ versus $Y = 3$, respectively.

Model (1) is implemented in the `gamlss` package [3] in R software [2] and can be accessed using the `MN3()` function. Setting logarithmic link functions for $g_1(\cdot)$ and $g_2(\cdot)$, in (1) we get

$$\begin{aligned} P(\boldsymbol{Y} = \boldsymbol{1}) &= \frac{\exp(\boldsymbol{X}_1\boldsymbol{\beta}_1)}{1 + \exp(\boldsymbol{X}_1\boldsymbol{\beta}_1) + \exp(\boldsymbol{X}_2\boldsymbol{\beta}_2)}, \\ P(\boldsymbol{Y} = \boldsymbol{2}) &= \frac{\exp(\boldsymbol{X}_2\boldsymbol{\beta}_2)}{1 + \exp(\boldsymbol{X}_1\boldsymbol{\beta}_1) + \exp(\boldsymbol{X}_2\boldsymbol{\beta}_2)} \quad \text{and} \\ P(\boldsymbol{Y} = \boldsymbol{3}) &= \frac{1}{1 + \exp(\boldsymbol{X}_1\boldsymbol{\beta}_1) + \exp(\boldsymbol{X}_2\boldsymbol{\beta}_2)}. \end{aligned} \tag{2}$$

in which $\boldsymbol{X}_k$ is a known model matrix of order $n \times (m_k + 1)$ and m_k denotes the number of explanatory variables related to the k^{th} parameter.

In order to measure how accurate the predictions are using the fitted model estimates from the training data set, we use the 0-1 loss function (or misclassification error). Using the validation data set, the precision of the estimated model is obtained by calculating

$$E = \frac{1}{v}\sum_{i=1}^{v} |sgn(y_i - \hat{y}_i)|, \tag{3}$$

where $v = n - \xi$ is the length of the validation data set and sgn is the signum function. The function E returns the proportion of errors for the $\hat{Y}$ of the validation data set, based on the model obtained using the training data set. The total sample sizes considered for the training data set and validation data set in this paper are 2/3 and 1/3 of the total sample size N, respectively.

3 RESULTS AND DISCUSSIONS

In order to construct the regression model, we are using a training set composed by $n_t = 5,868$, of which $n_{t1} = 1,270$, $n_{t2} = 2,505$ and $n_{t3} = 2,093$, where n_{ti} is the sample size for training set for response i, totalizing 2/3 of the total sample $n = 8,801$ containing observations for the response variable. The random samples were selected randomly proportionally (considering 1/3 for each level of Y), using the `sample` function in R software [2]. Using the *stepAIC.ALL* method to select additive terms for the different parameters of the MN3 distribution, we obtained the following final model

$$\begin{aligned} &\mu = \exp(\beta_{01} + \beta_{11}NIR + \beta_{21}RE + \beta_{31}NIR \times RE) \qquad \text{and} \\ &\sigma = \exp(\beta_{02} + \beta_{12}NIR + \beta_{22}RE + \beta_{32}G + \beta_{42}R + \beta_{52}RE \times G + \beta_{62}NIR \times G). \end{aligned} \tag{4}$$

Table 1 provides the maximum likelihood estimates (MLEs), standard errors (SEs) and p-values obtained from the fitted MN3 model based on the GAMLSS framework. All parameters are significant at the 5% significance level, indicating the accuracy of the method to select the additive terms. The results in this table indicate that the color spectrum NIR, RE, and the interaction $NIR \times RE$ are significant factors to model the odds $P(Y=1)/P(Y=3)$, represented by μ. Also, the color spectrum NIR, RE, G, R, and the interactions $RE \times G$ and $NIR \times G$ are influent factors to model the odds $P(Y=2)/P(Y=3)$, given by σ parameter.

Table 1: The MLEs, corresponding SEs and p-values of the estimates from the fitted MN3 model based on the GAMLSS framework.

Parameter	Estimate	SE	p-value	Parameter	Estimate	SE	p-value
β_{01}	-28.248	3.812	<0.001	β_{02}	-83.175	5.557	<0.001
β_{11}	0.412	0.057	<0.001	β_{12}	0.857	0.037	<0.001
β_{21}	3.233	0.172	<0.001	β_{22}	-1.845	0.152	<0.001
β_{31}	-0.034	0.002	<0.001	β_{32}	1.172	0.135	<0.001
				β_{42}	-0.059	0.018	0.001
β_{62}	-0.013	0.001	<0.001	β_{52}	0.033	0.004	<0.001

Based on the fitted models for μ and σ, Figure 1 displays the estimated probabilities (2) as functions of the interactions selected in (4). As we can see in Panel (a), considering the average of G and R, the smaller are the levels of NIR and RE, the greater is the probability of Y to be soil, the greater are the levels of NIR and RE the greater is the probability to be sugar cane and for moderate values of NIR and $RE > 20$ there is a high chance of Y being weed. Considering the average of NIR and R, Panel (b) shows that low values of RE and high values of G results in a high probability to be soil, while medium values for both RE and G will result in a high probability of Y to be sugar cane and low values for G and high values for RE is probably weed. Finally, considering the interaction between NIR and G (Panel (c)) we can notice that when the values of G are combined with low values of NIR the probability is high to be soil, when values of G are combined with high values of NIR there is a high probability to be sugar cane and when we combine medium values of both G and NIR there is a high probability to be weed.

To conduce the model validation, we are using the validation set composed by $n_{v1} = 635$, $n_{v2} = 1,252$ and $n_{v3} = 1,046$, where n_{vi} is the sample size for the validation set for $y = i$, totalizing 1/3 of complete sample $n = 8801$. Using equation (3) to calculate the proportion of error, based in the proposed model (4), we obtained $E = 0.031$, that is, the accuracy $(1 - E)$ to predict new responses based in the explanatory variables is 96.9%, which is an extreme high percentage confirming once again the great fit to the data provided by the MN3 distribution model based on the GAMLSS framework.

Now, using the fitted model (4), we predict 119,052 missing values of Y. The complete responses of length $N = 127,853$, composed by Y and $\hat{Y}$, are $n_1 = 20,449$, $n_2 = 72,853$ and $n_3 = 34,551$, where n_i represent the sample size for response $Y = i$. The results reveal that the level of weed infestation, considering both fields is 27.02%. A visual representation of the weed invasion is reported in Figure 2.

4 CONCLUSION

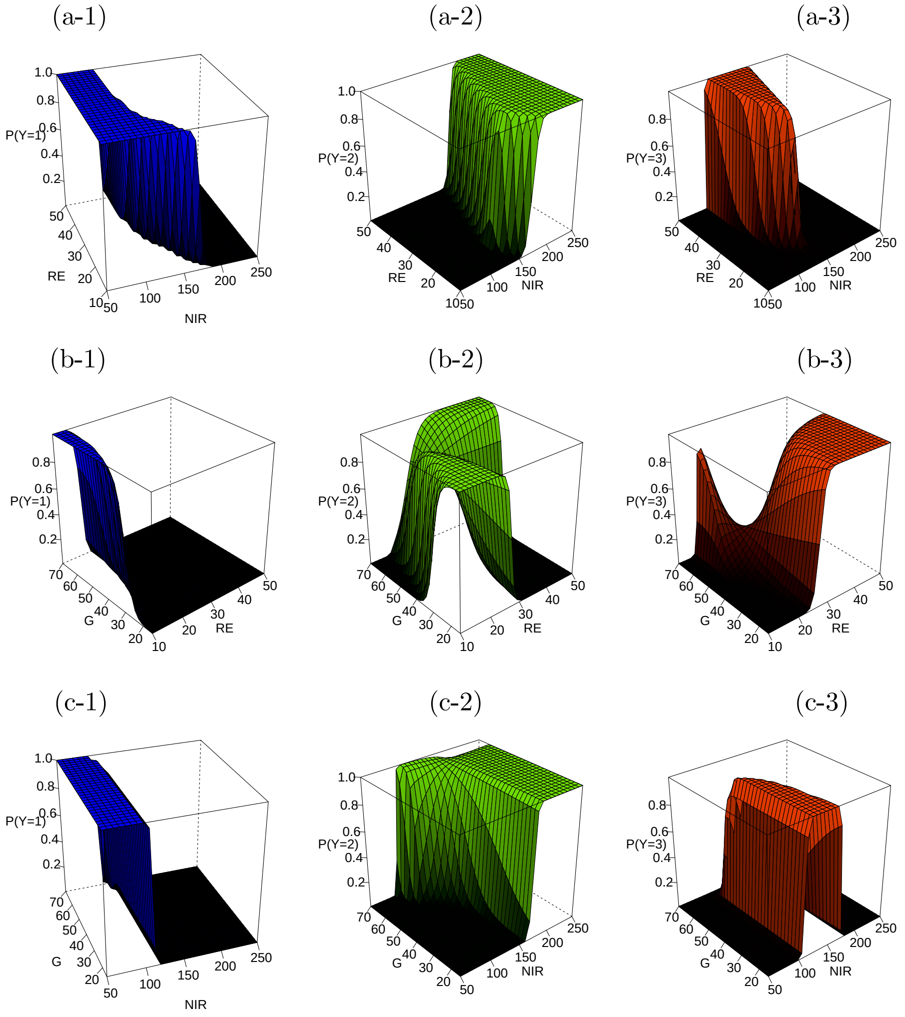
In this paper we use a multinomial logistic regression under the GAMLSS framework, in order to predict weed infestation in a sugarcane cultivar. The proposed model presented a 96.9% prediction power rate. Using the proposed model, it was possible to predict 119,052 missing values based only in the color information obtained directly from the field.

References

[1] Hosmer D.W.; Lemeshow, S. *Applied Logistic Regression*, 2nd ed., Wiley-Interscience Publication, New York, 2000.

[2] R Core Team, *R: A language and environment for statistical computing.* Software available at https://cran.r-project.org.

[3] Stasinopoulos D.M.; Rigby, R.A. *Generalized additive models for location, scale and shape (GAMLSS) in R*, Journal of Statistical Software 23 (2007), pp. 1–10.


(a-1) (a-2) (a-3)

(b-1) (b-2) (b-3)

(c-1) (c-2) (c-3)

Figure 1: (a) The estimated probabilities as functions of (a) RE and NIR, (b) G and RE and (c) G and NIR, considering the average for the fixed variables, for (1) $P(Y = 1)$, (2) $P(Y = 2)$ and (3) $P(Y = 3)$

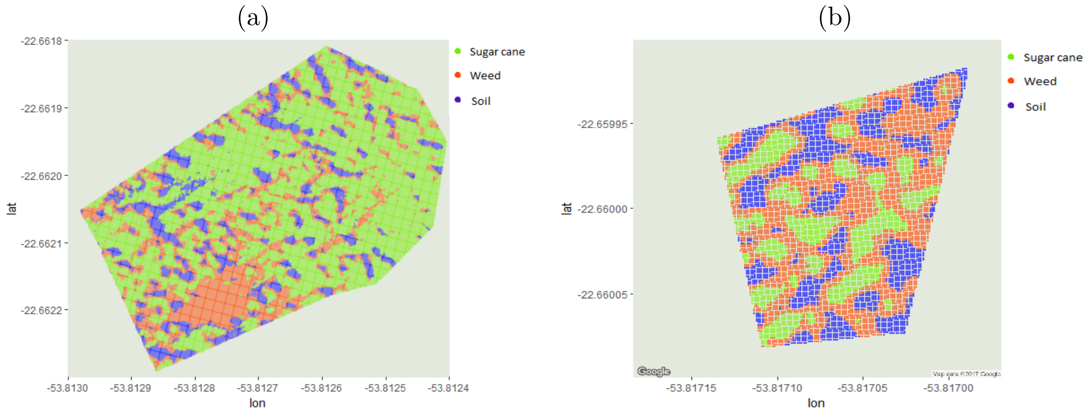

(a) (b)

Figure 2: Observed and predicted values of the response variable Y for: (1) field 1 and (b) field 2.