Sugarcane Land Classification with Satellite Imagery using Logistic Regression Model

F Henry[1], D E Herwindiati[2], S Mulyono[3], and J Hendryli[4]

[1,2,4] Faculty of Information Technology, Tarumanagara University, Letjen S. Parman No. 1, Jakarta, Indonesia

[3] The Indonesian Agency for the Assessment and Application of Technology,, Kawasan Puspitek, Serpong, Tangerang, Indonesia

E-mail: 535120072@fti.untar.ac.id, dyahh@fti.untar.ac.id, sidik.mulyono@bppt.go.id, jansonh@fti.untar.ac.id

Abstract. This paper discusses the classification of sugarcane plantation area from Landsat-8 satellite imagery. The classification process uses binary logistic regression method with time series data of normalized difference vegetation index as input. The process is divided into two steps: training and classification. The purpose of training step is to identify the best parameter of the regression model using gradient descent algorithm. The best fit of the model can be utilized to classify sugarcane and non-sugarcane area. The experiment shows high accuracy and successfully maps the sugarcane plantation area which obtained best result of Cohen's Kappa value 0.7833 (strong) with 89.167% accuracy.

1. Introduction

Indonesia is one of the countries in the world with great potential in the agricultural sector, which is important in the development of the country's economy. Crops plantation will have a high economic value if managed well. Monitoring agriculture process is one of the ways to manage crops plantation that can be done using satellite imagery data [1]. Utilizing satellite data has many benefits, such as obtaining information on the area that is difficult to reach like forests, swamps, and mountains which can help in agricultural planning such as irrigation [2], growth phase monitoring [3], and crop yield estimation [4]. Remote sensing technology also makes it easier and cheaper to map a very large area.

Sugarcane is one of the basic materials in sugar production and the increase in demand in Indonesia results in a greater need for suitable land for the plantation. Indonesia has a large sugarcane plantation area with more than 1.3 million farmers families involved and 2.5 to 3.0 million metric tons annual sugar production potential [5]. This paper discusses the mapping of sugarcane plantation from Landsat-8 satellite imagery. The process is broken down into two steps: training and classification. Polygon data, surveyed by the Indonesian Agency for the Assessment and Application of Technology (BPPT), are used as input for the training stage. The reflectance of sugarcane is selected according to the known wavelengths of sugarcane vegetation to calculate the normalized difference vegetation index (NDVI) using the NIR and red band reflectance values. After obtaining the set of data, multivariate analysis is used to build the logistic regression model to detect the sugarcane plantation in another area.

2. Related Works

As early as 1991, Lee-Lovick and Kirchner [6] had shown that sugarcane canopies give important spectral information which can help the task of classification from satellite imagery. The relationship between spectral reflectance and sugarcane canopy architecture had also been reported by [7]. As [8] reported, the spectral response of sugarcane is also influenced by chlorophyll, carotene, and other various pigments. Therefore, several vegetation indices, such as normalized difference vegetation index (NDVI) [9] and enhanced vegetation index (EVI) [10], are usually utilized as features to recognize the sugarcane plantation. A study from [11] shows a high relationship between MODIS NDVI and rainfall patterns on the sugarcane plantation.

Numerous models had been proposed in previous researches. Unsupervised classifier showed promising result of detecting sugarcane plantation from MODIS satellite imagery and EVI features in [10]. Another study from [12] used k-means clustering to classify sugarcane area from EVI temporal data after noise reduction using wavelet analysis. Lee-Lovick and Kirchner [6] studied the classification from Landsat TM imagery and found that band 1, 2, and 3 spectral reflectances are useful for identification of sugarcane. A study from [13] showed a model to discriminate sugarcane from coffee and citrus plantation in Brazil with a high accuracy of 95%.

3. Logistic Regression

Logistic regression is a mathematical model approach that can be used to describe the relationship of several independent variables on a dichotomous dependent variable [14]. In this paper, the model parameter (x) is the time series of NDVI value and the class variable (y) is state of sugarcane class (1) or non sugarcane class (0). in the case of sugarcane land classification. The logistic regression model predicts the class through the probability calculated from the sigmoid function $g(x)$.

$$h_\theta(x) = g(\theta^T x) = \frac{1}{1 + e^{-\theta^T x}} \tag{1}$$

where the range value is $0 \leq h_\theta(x) \leq 1$. The target y can determine from the value of h_θ as in Equation 2.

$$y = \begin{cases} 1 & h_\theta(x) > 0.5 \\ 0 & h_\theta(x) \leq 0.5 \end{cases} \tag{2}$$

The regression coefficient θ is estimated using maximum likelihood estimation (MLE). The MLE of logistic regression cannot be expressed in a closed-form, so an iterative method is used to maximized the likelihood function. We use gradient descent algorithm to minimize the cost function $J(\theta)$.

$$J(\theta) = -\frac{1}{m} \sum_{i=1}^{m} y^{(i)} \log h_\theta(x^{(i)}) + (1 - y^{(i)}) \log \left(1 - h_\theta(x^{(i)})\right) \tag{3}$$

where m is the number of training samples, $y^{(i)}$ is the target, and $x^{(i)}$ are the features of each data in the training set. The gradient of the cost function is defined as

$$\frac{\partial}{\partial \theta_j} J(\theta) = \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)}\right) x_j^{(i)} \tag{4}$$

and at each iteration in the gradient descent updates the parameter $\theta_{new} = \theta_{old} - \alpha \frac{\partial}{\partial \theta_j} J(\theta)$ where α is the learning rate.

The completed algorithm of the training step is as follows:

(i) Choose the initial value for θ, learning rate α, and error rate ϵ.
(ii) Repeat the following steps until the error converges $|J(\theta_{new}) - J(\theta_{old})| > \epsilon$.
  (a) Calculate the new sigmoid function h_θ and $\frac{\partial}{\partial \theta_j}$.
  (b) Simultaneously update the parameter θ.
(iii) The parameter θ after the iterations is the coefficient of logistic regression model.

4. Experiments and Results

4.1. Data

The input to the model is the series of normalized difference vegetation index (NDVI) for sugarcane and non-sugarcane land, consisting of 17 periods in a year. These periodic data follow the one-year phenology of sugarcane. The spectral values are obtained using ENVI 5.1 software.

The NDVI is a numerical indicator that uses a visible band and near-infrared of the electromagnetic spectrum and is widely adopted in the analysis of green vegetation in remote sensing. The formulation of NDVI is described as

$$\text{NDVI} = \frac{NIR - RED}{NIR + RED} \tag{5}$$

where NIR is the value of near-infrared spectral reflectance and RED is the value of visible spectral reflectance. The NDVI-series will be formatted and combined sequentially by the date of capture through the layer stacking process and served as the input in the training step.

The aim of the training step is to find the best logistic regression model to classify the sugarcane plantation. It will predict the sugarcane and non-sugarcane area through the logistic probability model, calculated from the sigmoid function in Equation 1 and 2. Meanwhile, the testing process is done to test the accuracy of the sugarcane area classification using logistic regression. We have 1800 locations of sugarcane and non-sugarcane area, which are employed in these two experiments. The first scenario utilizes 60% of it as the training data and the second scenario uses 25%, as the rest are for testing the trained model. These locations are from 17 satellite images of Kawadenan, East Java from May 2014 through May 2015.

4.2. Experimental Settings

The preprocessing of the Landsat-8 satellite imagery data is largely done using the ENVI software. For the classification process, we develop classification software using .NET framework 4.5. Our initial logistic regression parameters are initialized to be 0 and the learning rate is set to 0.5. We run the model training for 800 iterations. In every hundredth iteration, we save the parameters and calculate the performance.

4.3. Evaluation Measure

We evaluate the performance of the logistic regression model for sugarcane plantation area classification by calculating the prediction accuracy, Cohen's kappa statistic, and pseudo-R^2. Cohen's kappa statistic, first introduced by [15] can be interpreted as a interclass correlation coefficient [16]. It can also be used to measure the interrater reliability [17]. Cohen's kappa coefficient κ can be defined as follows:

$$\kappa = \frac{OA - AC}{1 - AC} \tag{6}$$

where OA is the sum of true positive and true negative divided by the total number of data and AC can be calculated as in Equation 7.

Table 1. The interpretation of the Cohen's kappa coefficient

Value of κ	**Strength of Agreement**
< 0.20	Poor
$0.21 - 0.40$	Fair
$0.41 - 0.60$	Moderate
$0.61 - 0.80$	Good
> 0.80	Very Good

$$AC = \left(\left(\frac{P}{n}\right) * \left(\frac{P'}{n}\right)\right) + \left(\left(\frac{N}{n}\right) * \left(\frac{N'}{n}\right)\right) \tag{7}$$

where P is the sum of true positive and false positive, N is the sum of false negative and true negative, and P' is the sum of true positive and false negative. The Cohen's kappa coefficient shows the strength of agreement between two variables as in Table 1 [18].

To evaluate the goodness-of-fit of the logistic model, some pseudo-R^2 measures have been developed, such as Efron's [19] and McFadden's [20] pseudo-R^2. These are "pseudo" measures because they look like R^2 in a sense that they are on the same scale. The Efron's pseudo-R^2 can be calculated as in Equation 8 with y_i is the target and $\bar{y}$ is the mean of the target data.

$$\rho^2 = \frac{\sum_{i=1}^{n} \left(y_i - h_\theta(x_{(i)})\right)^2}{\sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2} \tag{8}$$

Meanwhile, McFadden's pseudo-R^2 is defined as in Equation 9 where $LL(B)$ is log-likelihood of the full model and $LL(0)$ is log-likelihood of the null model. While the range of ρ^2 is similar to R^2 in regular regression analysis, yet the measurement cannot be determined with a standard index of R^2 and standard of "good fit" [20].

$$\rho^2 = 1 - \frac{LL(B)}{LL(0)} \tag{9}$$

4.4. Results

Figure 1 shows the pseudo-R^2 for each iteration in the training stage. The figure shows that the best model is acquired in 500 iterations. Although this is the case, we found that in the testing stage performance of the model from 600 iterations gives the best prediction accuracy. Table 2 shows the best Cohen's kappa coefficient and accuracy from the model.

It is interesting to see that with only 25% data as the training input, the trained logistic model can identify the sugarcane plantation area with more than 85% accuracy. The Cohen's kappa coefficient also shows that the model is good. Confusion matrix in Table 3 further shows the performance of this model on 25% proportion of the data.

In Figure 2 we show the mapping of sugarcane plantation in Kawadenan region, East Java. On the left is the ground truth data obtained by the BPPT and on the right is the classification result from the logistic regression model. The green area shows the sugarcane plantation area detected by the model, while the red area is the sugarcane real plantation. The logistic regression model detected much of the plantation area, but some misclassifications are observed.

5. Conclusion and Future Works

We have presented the logistic regression model for classifying sugarcane plantation area in Kawadenan region, East Java. The parameters learned in 500 iterations show the best pseudo-R^2 but in the testing stage, the prediction accuracy is lower than the parameters obtained from

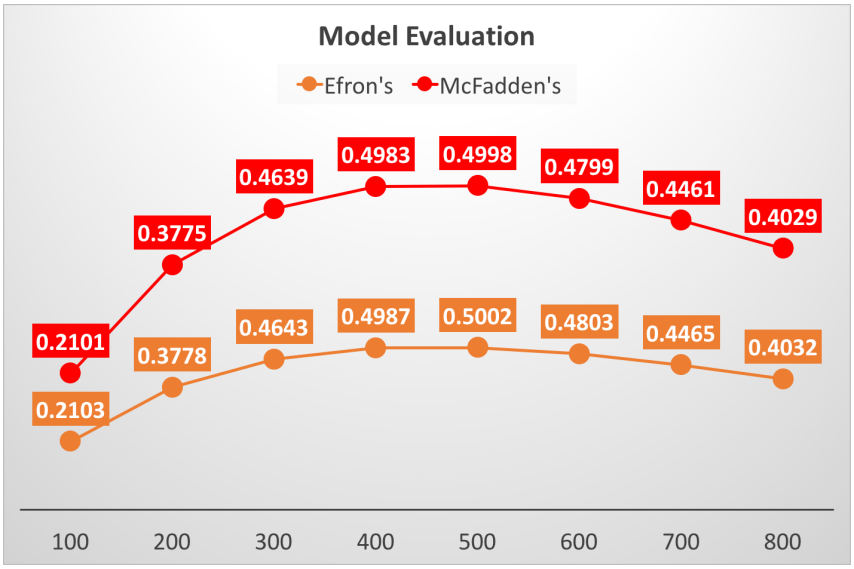

Figure 1. Pseudo-R^2 from our model training

Table 2. Cohen's kappa coefficient and accuracy from our experiments

		Proportion	
		60% training	**25% training**
Cohen's Kappa	Training	0.6 (good)	0.6933 (good)
	Testing	0.7833 (good)	0.7141 (good)
Accuracy	Training	80%	84.667%
	Testing	89.167%	85.704%

Table 3. Confusion matrix from the model trained only with 25% proportion of the data

		Prediction	
		Sugarcane	Non-sugarcane
Actual	Sugarcane	542	60
	Non-sugarcane	133	615

600 iterations. The model from 600 iterations predicted the sugarcane plantation with more than 85% accuracy even with only 25% of 1800 ground truth data. For the future works, we plan to utilize better learning algorithm, such as stochastic gradient descent to help to minimize the cost function of logistic regression faster and more effective. More features to describe the sugarcane plantation can also be explored further to increase the prediction accuracy.

References

[1] Frolking S, Qiu J, Boles S, Xiao X, Liu J, Zhuang Y, Li C and Qin X 2002 *Global Biogeochemical Cycles* **16**
[2] Kim Y, Evans R G and Iversen W M 2008 *IEEE Trans. on Instrumentation and Measurement* **57** 1379–1387
[3] Moran M S, Inoue Y and Barnes E 1997 *Remote sensing of Environment* **61** 319–346
[4] Shanahan J F, Schepers J S, Francis D D, Varvel G E, Wilhelm W W, Tringe J M, Schlemmer M R and Major D J 2001 *Agronomy Journal* **93** 583–589
[5] Wright T and Meylinah S 2016 Indonesia sugar annual report 2016 Tech. rep. U.S. Department of Agriculture Foreign Agricultural Service
[6] Lee-Lovick G and Kirchner L 1991 *Proc. the Australian Society of Sugar Cane Technology* vol 13 pp 124–129

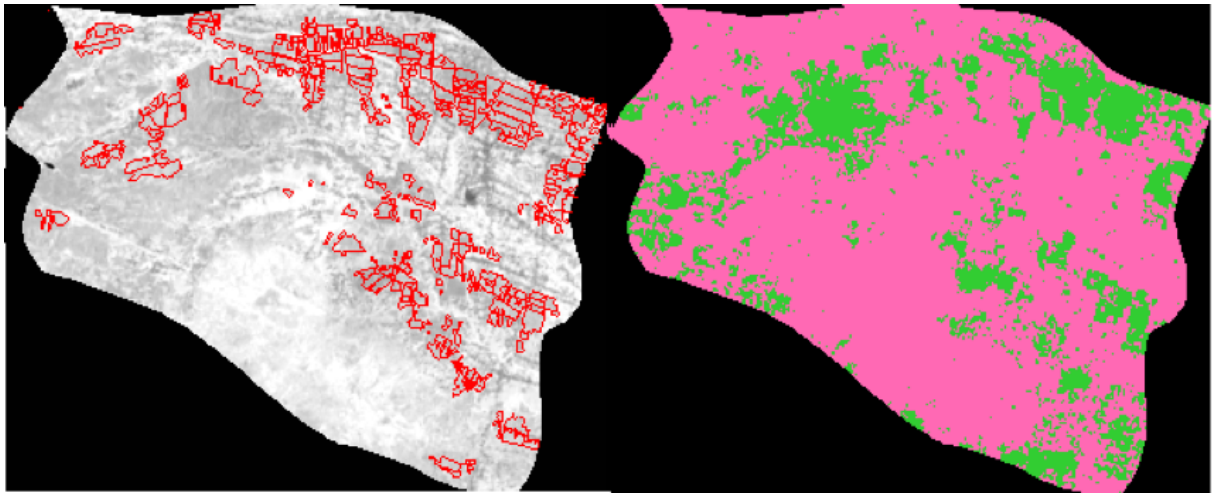

Figure 2. Mapping of ground truth sugarcane plantation area (left) and logistic regression model prediction (right)

[7] Galvao L S, Formaggio A R and Tisot D A 2005 *Remote Sensing of Environment* **94** 523–534
[8] van der Meer F, De Jong S and Bakker W 2002 *Imaging spectrometry* (Springer) pp 17–61
[9] Rahman M R, Islam A and Rahman M A 2004 *Plan Plus* **1** 1–12
[10] Xavier A C, Rudorff B F, Shimabukuro Y E, Berka L M S and Moreira M A 2006 *Int. J. Remote Sensing* **27** 755–768
[11] Gunnula W, Kosittrakun M, Righetti T L, Weerathaworn P and Prabpan M 2011 *Australian J. Crop Science* **5** 1845
[12] Rudorff B, Adami M, Aguiar D A, Gusso A, Silva W F and De Freitas R M 2009 *IGARSS (4)* pp 252–255
[13] Tardin A, Deassuncao G and Soares J 1992 *Pesquisa Agropecuaria Brasileira* **27** 1355–1361
[14] David K and Mitchel K 2010 *Logistic regression: A self learning text* 3rd ed (New York: Springer–Verlag Inc)
[15] Kohen J 1960 *Educ Psychol Meas* **20** 37–46
[16] Ebel R L 1951 *Psychometrika* **16** 407–424
[17] Smeeton N C 1985 *Biometrics* **41** 795
[18] Azen R and Walker C M 2011 *Categorical data analysis for the behavioral and social sciences* (Routledge)
[19] Scott Long J 1997 *Advanced quantitative techniques in the social sciences* **7**
[20] McFadden D, Talvitie A, Cosslett S, Hasan I, Johnson M, Reid F and Train K 1977 *Urban Travel Demand Forecasting Project, Phase* **1**