International Journal on Advanced Science Engineering Information Technology

Vol.12 (2022) No. 5
ISSN: 2088-5334

Data-driven Exploratory Analysis for Raster Data Using Self-Organizing Maps Regressor

Aulia Khoirunnisa Fajri [a,*], Indra Ranggadara [a], Suhendra [a], Aries Suharso [b]

[a] *Faculty of Computer Science, Mercu Buana University, Jakarta, Indonesia*
[b] *Faculty of Computer Science, Singaperbangsa University, Karawang, Indonesia*
*Corresponding author: *ahulemos@gmail.com*

***Abstract*— Sugarcane is one of the plantation commodities in Indonesia which has a big potential. Sugarcane growth consists of 4 phases that happen in a year. In the Grand Growth phase, sugarcane needs an appropriate condition to grow well and enter the next phase. The factors that affect sugarcane's Grand Growth phase are water, temperature, and sunlight. Rainfall is one of the sugarcane water sources needed, but the rainfall intensity is different, and the rainfall distribution is uneven every year. The uneven rainfall caused water stress in a sugarcane plantation. That is why it is necessary to identify the water content in sugarcane plantations to maintain the quality of sugarcane. This study predicted the water content of sugarcane plantations so the areas indicated with water stress can be anticipated. Raster data are collected from Landsat-8 satellite imagery and analyzed using one of the data-driven exploration analysis methods, PCA (Principal Component Analysis), to analyze the overlay of the Landsat 8 imageries of the sugarcane plantation area. After that, the raster data were processed to calculate the water index of the sugarcane plantation, known as NDWI (Normalized Different Water Index). NDWI values of the sugarcane plantation area are converted into an array and then become data input for the Self-Organizing Map Regressor algorithm to predict the water content of the sugarcane plantation. The results are predicted water index values for the sugarcane plantation with 72% accuracy.**

***Keywords*— NDWI; PCA; self-organizing map; sugarcane; water stress.**

Manuscript received 6 Dec. 2020; revised 22 Jun. 2021; accepted 4 Aug. 2021. Date of publication 31 Oct. 2022.

I. INTRODUCTION

With the rapidly developing computing technology, data-driven machine learning methods have become increasingly popular in the last decade in all fields related to data and modeling, including plantations. A model with a data-driven approach is based on the observed relationship between input and output variables. With a variety of learning algorithms, the data-driven approach provides a flexible way to model natural phenomena such as water stress. A data-driven approach is suitable for use where: 1) current human insights into the process are insufficient to provide an explicit modeling strategy, or 2) physical modeling is too challenging due to complex target processes [1].

Sugarcane is a commodity that plays a vital role in the economy of Indonesia. Indonesia's average sugarcane plantation area from 2015 to 2018 was 400,000 hectares [2]. This makes sugarcane a source of income for thousands of farmers in the Indonesian plantation industry. Sugarcane is also a relatively cheap source of calories for the people of Indonesia.

Sugarcane plantations in Indonesia are divided into large plantations and smallholder plantations according to their concessions. Large plantations consist of large state plantations and large private plantations. In Indonesia, the most massive sugar production is produced by smallholders, with a percentage of 55%[2].

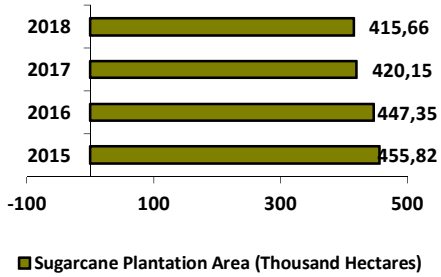

Fig. 1 Sugarcane plantation area in Indonesia 2015 - 2018

Figure 1 shows the sugarcane plantation area in Indonesia from 2015 to 2018. In 2015, the area reached 455,820 hectares. In 2016, there were 447,350 hectares of sugarcane plantation. In 2017, the area of sugarcane plantations was 420,150 hectares wide, and in the following year, in 2018, the area reached 415,660 hectares. The chart shows that the average area of sugarcane plantations in Indonesia between 2015 and 2018 is 434,745 hectares.

Sugarcane has four growth phases: Crop Establishment, Tillering, Grand Growth, and Maturity & Ripening. In the Crop Establishment phase, sugarcane plants are still in the form of seeds taken from sugarcane stalks with 2-3 buds that have not yet grown. In Tillering phase, the leaves begin to bloom, and new roots emerge from the base of the shoot. In the Grand Growth phase, leaf crowns begin to appear, roots begin to develop, and the plant grows 3-4 stems per month. This phase is followed by the Maturity and Ripening phase, where the sugarcane stalks begin to fill the growth of new leaves and internodes slow down.

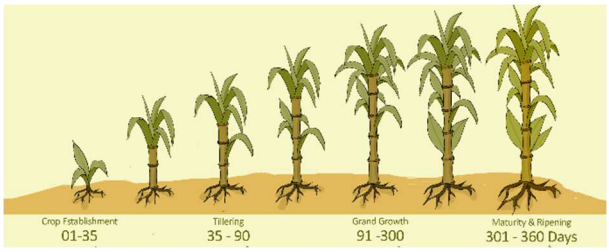

Fig. 2 Sugarcane growth phase

Figure 2 shows the sugarcane growth phase followed by the duration of each phase [3]. The Crop Establishment phase takes place for 3 until 35 days. The second phase is Tillering phase which occurs from the 35th days after planting until the 3rd month. The Grand Growth phase happens from the 3rd month after planting until the 9th month. After 9 months, the sugarcane plant enters Maturity and Ripening phase that can occur

In the Grand Growth phase, sugarcane requires more water than in other phases for stem elongation and sugar production [4]. Rain is one of the water sources for sugarcane, but rainfall has a varying amount and distribution every year, so we need a way to anticipate fluctuating rainfall. One way is to make irrigation channels [5]. Irrigation channels should be installed in plantation areas that have the potential to experience water stress. Predictions can be made using data-driven analysis to find out the potential areas.

Data-driven exploratory analysis is a data-driven analysis that uses large amounts of data and aims to gain insight and examine further data by identifying patterns in the data [6], [7]. This analytical method has been popularly used in the last decade to investigate phenomena in various fields using various kinds of datasets. To predict water stress using data-driven analysis, we can use raster data.

Raster data is a data model in geographic information systems to describe and represent aspects of the real world to a computer [8]. A raster data model consists of a collection of rows and columns containing pixels of the same size and linked together. The raster data model is also known as a grid-based system. The raster data model averages all the values in a pixel to produce a single value. The larger the area represented by a pixel, the less accurate the data. The extent of the area represented by a pixel determines the spatial resolution of the original raster model. The resolution is determined by the size of one side of the pixel. For example, a raster model with pixels representing 100 meters (10 m x 10 m) in the real world can be said to have a spatial resolution of 10 m [8]. The raster data model's advantage is the technology needed to make raster images affordable and can be found anywhere, for example, in digital cameras and cellphone cameras. Several major satellites also continuously transmit the latest raster graphics to scientific facilities worldwide, and some of these facilities make available images from these satellites free of charge. Another advantage is that raster images have a simple data structure.

Each pixel has a value that describes a spatial phenomenon's characteristics in a location according to its rows and columns. The value can be decimal or floating-point and was processed for data-based analysis and become the basis for predicting water stress in sugarcane plantations.

The emergence of big data has caused the development of artificial neural networks (ANN) as a tool for classification and regression. The model used varies depending on the work performed and the available data. The types of ANN currently popular in remote imaging research are feed-forward neural networks and convolutional neural networks. Both types of ANN require a large amount of training data. There are conditions where the reference data (ground truth) is limited to applications because the acquisition takes a long time and costs a lot. Data reference limitations can be on the amount, accuracy, and quality, so machine learning approaches that rely on large amounts of training data are often not applicable. SOM is a type of ANN that can address limited reference datasets [9].

Self-Organizing Map or SOM is an artificial neural network that can process datasets with few references, such as raster data. SOM was founded by Teuvo Kohonen in 1982 and came into wide use in 1990. Kohonen described SOM as “an analysis and visualization tool for high-dimensional data”. SOM has a characteristic similar to the human brain [10].

An ANN architecture SOM consists of two connected layers: the input layer and the two-dimensional grid output layer. Neurons on the output layer are interconnected to each other using a neighborhood relationship. This attribute of SOM lowers the overfitting of the training data, and the 2D output layer visualizes the SOM completely. The SOM algorithm is an unsupervised learning method. In other words, it is an algorithm that learns without guidance [9]. The advantage of the Self-Organizing Map algorithm is that it is easy to implement and can solve nonlinear problems with very high complexity. Another advantage is that this algorithm can handle missing data, small data dimensions, and large sample or input data sizes [11].

SOM is popularly used for clustering, classification, data mining, and prediction [12]–[15] but there are also many studies which applied SOM for regression with raster data [9], [16], [17]. SOM performance in regression is better than supervised learning methods and other modeling methods [18], [19]. This research aims to analyze raster data using data-driven exploratory analysis and then use the SOM algorithm to predict the water stress of land in sugarcane plantations.

II. Materials and Method

A. Data

The data used in this study were obtained from Landsat-8 satellite imagery, downloaded from the USGS EarthExplorer page (URL: earthexplorer.usgs.gov). Landsat-8 imagery consists of 11 bands. For this research, the bands used from each image are band 5 and band 6, so that there are eight imageries in total.

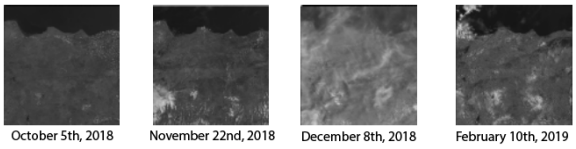

Fig. 3 Example of raster data from Landsat-8 satellite imagery

Figure 3 shows downloaded raster data from USGS EarthExplorer page. The data downloaded consists of 4 periods: October 5th, 2018, November 22nd, 2018, December 8th, 2018, and February 10th, 2019. The download results are in the form of 4 images, each of which consists of 11 bands. The eight imageries are raster data converted into arrays. The raster data then goes through the Clipping stage to be adjusted to the research location, one of the sugar cane plantation areas in Kediri, East Java. East Java is one of the sugarcane production centers in Indonesia.

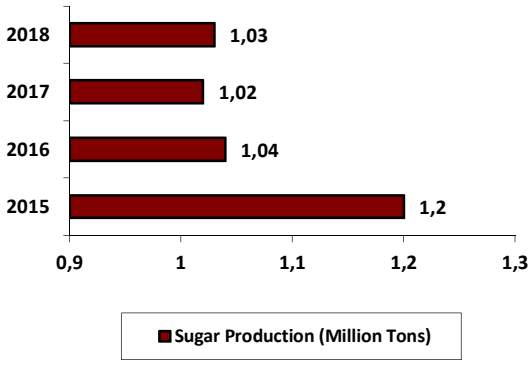

Fig. 4 The amount of sugar production in East Java, Indonesia

Figure 4 shows sugar production in East Java from 2015 to 2018. In 2015, the total sugar production in East Java reached 1.2 million tons. In 2016, the amount of sugar produced in East Java was 1.04 million tons. In 2017, there are 1.02 million tons produced in East Java, and in the following year, there are 1.03 million tons of sugar produced in East Java. We can see East Java's average sugar production is 1,07 million tons per year from the chart.

B. Self-Organizing Map

The SOM algorithm concept provides many classification resources (often referred to as neurons) organized based on the available classification patterns (also known as input patterns). Parts of the SOM can be activated with certain types of input patterns. The SOM is trained iteratively using a large number of epochs. Epoch is the processing of all input patterns once. Each input pattern was processed as many as the number of epochs [10]. SOM calculates the distance of each input data with each neuron. After that, the neuron with the closest distance to the input data is selected. The neuron is called the winning neuron, and then the input data is mapped to the winning neuron. Furthermore, the neuron was moved towards the position of the input pattern. This movement distance is controlled by a parameter known as the learning rate. It is necessary to correct the position of the winning neuron and the position of the surrounding neurons to maintain neuron connections in the output area [10]. The steps of SOM are described in detail as follows.

SOM has two stages. The first stage is training. A SOM map was generated from the input data at this stage. The second stage is the classification of the SOM map that has been made. There are three stages in the SOM training process: initialization, matching, and updating. At the initialization stage, the size of the grid is determined, which become the two-dimensional SOM map. Then we determine the weight vector for each neuron in the output layer. Next is to determine the learning rate and the number of iterations. Learning rate controlled the value of changes in the weight vector. The learning rate value ranges from 0 to 1. The learning rate value decreased as the iteration progresses. The learning rate value was determined randomly [11]. The number of iterations is also determined randomly, but to improve the accuracy of the SOM map, the number of iterations should be at least 500 times the number of neurons in the output layer [20]. The radius of the neighborhood is also determined to determine the area of the neighborhood. After that we enter the matching stage.

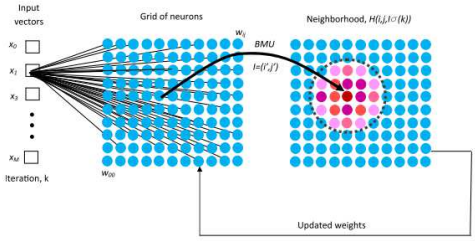

Fig. 5 The visualization of SOM unsupervised training.

Figure 5 shows visualization of matching stage and updating stage in SOM training. At matching stage, one input data is taken randomly, then the value of the data is used to calculate the distance between the input data and each neuron in the output layer. How to calculate the distance is by using the Euclidean distance formula, which can be seen in Equation 1, where x is the input data vector and wi is the weight of the input vector.

$$d_i(t) = \|x(t) - w_i(t)\| = \sqrt{\sum_{j=1}^{m}(x_{t\,j} - w_{t\,ji})^2} \quad i = 1, 2, \ldots\ldots, n \quad (1)$$

After the matching stage is complete, the winning neuron or neuron that has the smallest distance is determined from the input data. The winning neuron is denoted by c and is determined by the formula in Equation 2.

$$c(t) = \arg\min_i \{\|x(t) - w_i(t)\|\} \quad (2)$$

The winning neuron then moves closer to the input data. The next stage is the renewal stage, where the neuron weight vector is updated. The neuron weight vector is updated with Equation 3, where t is the discrete-time coordinate.

$$w_i(t+1) = w_i(t) + \alpha(t)[x(t) - w_i(t)] \quad (3)$$

It is necessary to consider the neighboring neuron information from the winning neurons to maintain the topological attributes of the input data in the output layer. The weight adaptation rate decreased in each iteration based on the neighborhood function hc, (t), where i is the index of neighboring neurons. The recommended neighborhood function to use is the Gaussian function as in Equation 4.

$$h_{c,i}(t) = \exp\left(-\frac{d_{ci}^2}{2\sigma^2(t)}\right) \quad (4)$$

where d_{ci}^2 is the distance between the winning neuron c and the neighboring neuron i, and σ^2 (t) is the neighborhood radius in the iteration t. After determining the neighbors of the winning neurons, the vector weights of the neighboring neurons are updated to Equation 5.

$$w_i(t+1) = w_i(t) + \alpha_i(t)h_{c,i}[x(t) - w_i(t)] \quad (5)$$

The second and third stages are repeated until the weight vector is similar to the input data or until it reaches the maximum number of iterations. The stages of the SOM training can be summarized into the following steps: 1) Create a SOM map; 2) Initialization of weight vector value from SOM map, learning rate, number of iterations, and neighborhood radius; 3) Select one input data; 4) Select the winning neuron by selecting the neuron with the closest distance to the input data; 5) Update the winning neuron weight and surrounding neurons; 6) Repeat steps 2-5 until the weight vector is similar to the input data. The visualization of SOM training can be seen in Figure 5.

The evaluation uses five parameters: Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), R^2, and Explained Variance Score. MAE and RMSE have been widely used in recommendation systems to assess the difference between the predicted results and the original value[21]. MAE is the average value of the number of errors in the prediction results of the SOM input data. The smaller the MAE value, the more similar the prediction results were with the input data, so is the RMSE value. The smaller the RMSE value, the more accurate the prediction results are [22]. The MSE value is used to show the best performance of the modelling to produce predictions. The R^2 value shows the correlation between the original data and the predicted results [23]. EVS determines the ratio between the number of data errors and the number of correct data.

C. SOM Regressor

SOM Regressor is a semi-supervised SOM method, combining unsupervised SOM and supervised SOM [24]. The unsupervised SOM section is applied first to the raster data. Then the supervised SOM section is applied afterward to the unsupervised SOM training data where there is less amount of data. Sample weights are applied to add weights to labeled data to be greater than unlabeled data. Class weighting only improved the SOM Regressor's performance if the labeled data is unbalanced in different classes.

Figure 6 shows the steps of semi-supervised SOM. First, the unsupervised SOM is applied, then the trained unsupervised SOM becomes data input for the supervised phase. The dimensions of weight and the training process of the two SOM methods above are different. The weights in the unsupervised SOM are on the same dimensions as the input data. By adapting these weights, the BMU changed for each data point. In contrast, the supervised SOM weight has the same dimensions as the regression target variable. The weights of the supervised SOM are one-dimensional and contain a continuous number. In other words, the unsupervised SOM is used to find BMU for each data point, while the supervised SOM links the selected BMU to a particular estimation.

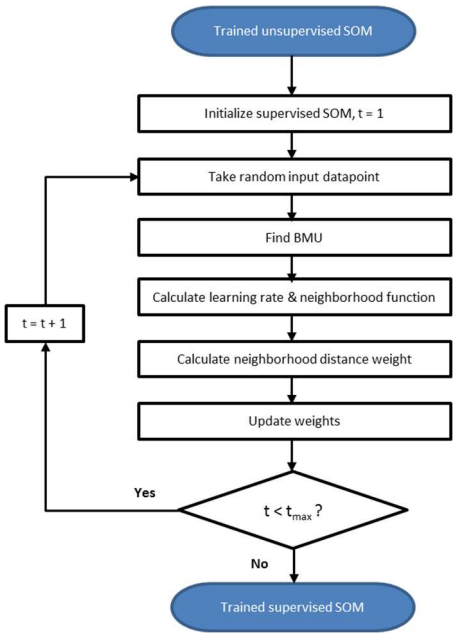

Fig. 6 Flowchart of semi-supervised SOM

D. Data-driven Exploratory Analysis

This study used data-driven exploratory analysis to see water content changes in the sugarcane plantation. Each pixel in collected raster data has a value that was processed to NDWI. NDWI values in each pixel was processed with Principal Component Analysis (PCA). PCA is a method for deriving dimensional data sets. According to Verbeeck et al., PCA can help reveal the dataset's structure, which is difficult to observe on high-dimensional datasets [25]. PCA describes a dataset with many variables that summarized the dataset's features [26] while maintaining good variation [25]. PCA is also used to explore similarities and hidden patterns [27] and detect changes in a dataset [28]. PCA has been applied in a wide range of applications in various fields ranging from biochemistry [29], tourism [30], geology [31], image processing [32], environment [33], and marine engineering [34]. PCA has been used to process raster data, as in the following studies [35], [36], [37].

Generally, PCA has six stages [38]. The first stage is the standardization of values in the input data. The purpose of this stage is so that the variables in the data can contribute equally. The second stage is to build a covariance matrix to determine the correlation or relationship between the variables in the input data. The next step is to calculate the eigenvalues where and eigenvector for each eigenvalue that was used to determine the main component.

Furthermore, the eigenvectors were sorted based on the highest eigenvalues to the lowest eigenvalues. The

eigenvector that has the highest eigenvalue was the eigenvector of the first principal component, and so on until the eigenvector that belongs to the lowest eigenvalue. We measure the amount of information possessed by one main component by determining the level of contribution. To get the level of contribution, we divide the eigenvalue of one principal component by the sum of all eigenvalues. The next step is to create a projection matrix or feature vector to determine the main components to be used. The feature vector is the result of dimensional degradation. For example, if we only want to use two eigenvectors on 3-dimensional data, the data dimension changed to 2. However, if the initial purpose of using PCA is to describe data with new variables without reducing the dimensions, it is better to use all existing eigenvectors. The last stage is to re-orient the original data to the data represented by the principal components by multiplying the original data transpose by the feature vector transpose.

This study used PCA to analyze water content in 4 data acquisition periods. We use NDWI or Normalized Different Water Index to determine the water content in the sugarcane plantation area. NDWI is an index used for remote sensing of water content in green vegetation from outer space. NDWI is obtained by processing the near-infrared (NIR) and shortwave infrared (SWIR) channels in Landsat 8 satellite imagery. NDWI helps determine which plantation areas have water bodies and which areas are dry. The reflection of SWIR light reflects changes in moisture content in vegetation and mesophyll structure in the vegetation canopy, while the reflection of NIR light is influenced by the leaves' internal structure and the degree of leaf dryness, not by the moisture content. The combination of NIR and SWIR channels eliminates variations caused by the internal structure of the leaves and the degree of leaf dryness, thereby increasing the accuracy in obtaining vegetation moisture [39]. The formula for obtaining NDWI is in Equation 9.

NDWI values range from -1 to 1, with an NDWI value above 0 indicating water level, while a value of 0 to less than or equal to 0 indicates a non-water surface [40]. The NDWI value that has been obtained is then analyzed using PCA. Figure 7 shows the steps taken in this research.

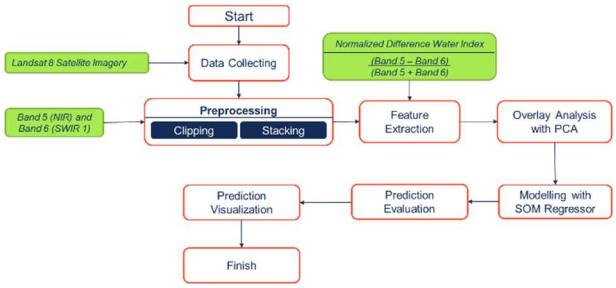

Fig. 7 Flowchart of the research method

III. Results and Discussion

This study conducted a data-driven analysis to determine water content changes in the sugarcane plantation area based on water index data from 4 data acquisition periods using PCA. The raster data is adjusted to the sugarcane plantation area and then processed with Equation 8 to determine the NDWI value of each image pixel.

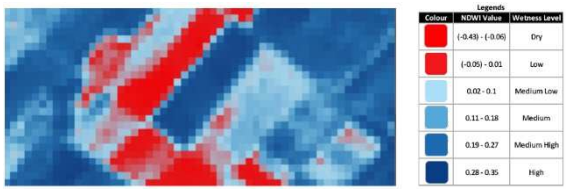

Fig. 8 Visualization of water content in sugarcane plantations before applying PCA

Figure 8 shows the classification of NDWI values. There are six classes that represent water content in sugarcane plantation: 1) Dry, 2) Low, 3) Medium Low, 4) Medium, 5) Medium High, 6) High. Before applying PCA, the highest value of NDWI in the sugarcane plantation is 0.35, and the lowest value is -0.43. NDWI values within the range -0.43 to -0.06 means no plants contain water or dry. NDWI values within the range -0.05 to 0.01 means low level of water. NDWI values within the range 0.02 to 0.1 shows that the area has medium to low level of water. NDWI values within the range 0.11 to 0.18 shows medium level of water content. NDWI values within the range 0.19 to 0.27 means the area has medium to high water level, and NDWI values within the range 0.28 to 0.35 show that the area has a high level of water content.

The NDWI value of the four images is then processed using PCA using the formula in Equations 5, 6, and 7 to determine the water content in the four image acquisition periods. The results of NDWI processing from raster data using PCA can be seen in Figure 8. These images are the Principal Components (PC) of water content in sugarcane plantation.

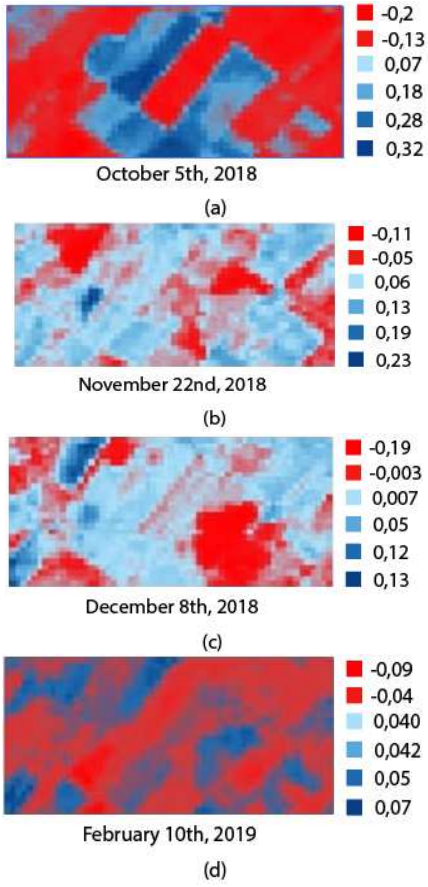

Fig. 9 4 PCs with their own NDWI class

In Figure 9, four images represent 4 PCs generated by PCA. In Figure 8a, the maximum value of the NDWI index for the period October 5th, 2018, was 0.32. In Figure 8b, the maximum value of the NDWI index for the period November 22nd, 2018, was 0.23. In Figure 8c, the maximum value for the NDWI index on December 8th, 2018, was 0.13, and in Figure 8d, the maximum value for the NDWI index on February 10th, 2019, was 0.07. After that, the classification is carried out based on the image's color representing the NDWI value. The red color shows an NDWI value of less than 0, which indicates that it is not a water body. The blue color indicates an NDWI value of more than 0, indicating a body of water.

In Figure 9a, it can be seen that the red blocks are more dominant than the red blocks. This may be due to the sugarcane plant being still in the Tillering phase, where the water is not the main supporting factor so that the water content in the sugarcane vegetation is not too high. In Figure 9b and 9c, it can be seen that the blue block dominates. This may be due to the sugarcane plant having entered the Grand Growth phase, where water is the main requirement so the water content in sugarcane plantations is relatively high. Then in Figure 9d, it can be seen that the red blocks fill the image again. This may be due to the sugarcane plant that has entered the Maturity and Ripening phase, where water content is not the main supporting factor in that phase. The NDWI index from the four image acquisition periods is combined to become input data in predicting water content in the sugarcane plantation area.

Prediction of water content in the sugarcane plantation area uses the SOM algorithm with Equations 1, 2, 3, 4, and 5. After predicting water content, the evaluation results are performed using the five previously mentioned parameters.

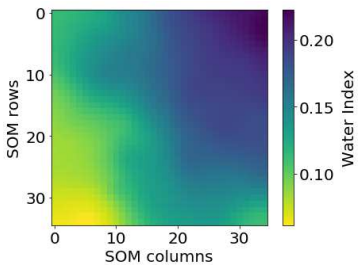

Fig. 10 SOM prediction visualization

Figure 10 shows the resulting SOM map. The map consists of 30 rows and 30 columns representing the predicted water index, with the value less than 0.10 in yellow color and the value more than 0.10 in blue shade. This research used cross-validation method to verified the accuracy of the modelling process in training stage [41]. The cross-validation method used is the 10-fold cross-validation method. The method divides the dataset into ten parts, and each part has one subset that is used as validation data and nine other subsets as training data. This 10-fold cross-validation method calculated the accuracy of each part (fold) to produce ten accuracy values [42]. The accuracy of this modeling is 72%. This figure is obtained from the average of 10 accuracy values in the second row generated by the 10-fold cross-validation method.

The MAE, RMSE, and MSE values that appear for the prediction results are 0.019, 0.025 and 0.00643. These values show that the prediction is accurate with low average of errors. Also, the test results show an R2 value of 79.2% and an EVS value of 0.62. Both of those values show that the correlation between the original data and prediction result is positive with low amount of error data.

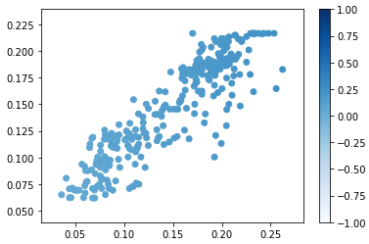

Fig. 11 The distribution of the NDWI value, which indicates the prediction of water content in the sugarcane plantation area

Figure 11 shows the distribution of NDWI in the prediction results where the index is dominated by a value of more than 0, which is the area of the water body on a sugarcane plantation with a maximum NDWI index predicted to reach 0.224.

IV. Conclusion

This study uses raster data from 4 periods to perform a data-driven explanatory analysis. The analysis results are used to see changes in water content and then become a reference for predicting water content in sugarcane plantations. We use the NDWI index to determine water content over four image acquisition periods, with an NDWI value below 0 indicating a non-water body area and an NDWI value above 0 indicating the area of a water body. The accuracy of the modeling process in the training stage is verified using the cross-validation method. The modeling results are then tested using five parameters: MAE, RMSE, MSE, R^2, and EVS, whose values indicate that the modeling has worked well with accurate results. Other supporting data can be used to increase the parameter value, especially the R2 value, such as rainfall data.

References

[1] W. Chang and X. Chen, "Monthly rainfall-runoffmodeling at watershed scale: A comparative study of data-driven and theory-driven approaches," *Water (Switzerland)*, vol. 10, no. 9, pp. 1–21, 2018, doi: 10.3390/w10091116.

[2] BPS, "Statistik Tebu Indonesia 2018," Badan Pusat Statistik RI, 2019.

[3] S. Marjayanti, "Teknik Budidaya Tebu," *Pelatih. Budid. Tanam. Tebu Pt Perkeb. Nusant. Xii*, pp. 1–21, 2012.

[4] M. S. de Camargo, B. K. L. Bezerra, A. C. Vitti, M. A. Silva, and A. L. Oliveira, "Silicon fertilization reduces the deleterious effects of water deficit in sugarcane," *J. Soil Sci. Plant Nutr.*, vol. 17, no. 1, pp. 99–111, 2017, doi: 10.4067/S0718-95162017005000008.

[5] A. S. Tayade, S. Vasantha, S. Anusha, R. Kumar, and G. Hemaprabha, "Irrigation Water Use Efficiency and Water Productivity of Commercial Sugarcane Hybrids Under Water-Limited Conditions," vol. 63, no. 1, pp. 125–132, 2020.

[6] W. Maass, J. Parsons, S. Purao, V. C. Storey, and C. Woo, "Data-driven meets theory-driven research in the era of big data: Opportunities and challenges for information systems research," *J. Assoc. Inf. Syst.*, vol. 19, no. 12, pp. 1253–1273, 2018, doi: 10.17705/1jais.00526.

[7] M. Komorowski, D. C. Marshall, J. D. Salciccioli, and Y. Crutain, "Secondary Analysis of Electronic Health Records," *Second. Anal. Electron. Heal. Rec.*, pp. 1–427, 2016, doi: 10.1007/978-3-319-43742-2.

[8] J. C. Campbell and M. Shin, *Geographic Information System Basics v.1.0*. 2012.

[9] S. Keller *et al.*, "Hyperspectral data and machine learning for estimating CDOM, chlorophyll a, diatoms, green algae and turbidity,"

Int. J. Environ. Res. Public Health, vol. 15, no. 9, pp. 1–15, 2018, doi: 10.3390/ijerph15091881.

[10] F. Bação and V. Lobo, "Introduction to Kohonen's Self-Organising Maps," *Inst. Super. Estat. E Gest. Inf.*, p. 22, 2010.

[11] U. Asan and S. Ercan, "An Introduction to Self-Organizing Maps," *Comput. Intell. Syst. Ind. Eng.*, vol. 6, no. March 2014, pp. 469–479, 2012, doi: 10.2991/978-94-91216-77-0.

[12] L. C. Chang, W. H. Wang, and F. J. Chang, "Explore training self-organizing map methods for clustering high-dimensional flood inundation maps," *J. Hydrol.*, vol. 595, p. 125655, 2021, doi: 10.1016/j.jhydrol.2020.125655.

[13] M. Milovanovic *et al.*, "A novel method for classification of wine based on organic acids," *Food Chem.*, vol. 284, no. January, pp. 296–302, 2019, doi: 10.1016/j.foodchem.2019.01.113.

[14] N. Chen, L. Chen, Y. Ma, and A. Chen, "Regional disaster risk assessment of china based on self-organizing map: Clustering, visualization and ranking," *Int. J. Disaster Risk Reduct.*, vol. 33, pp. 196–206, 2019, doi: 10.1016/j.ijdrr.2018.10.005.

[15] B. R. Shivakumar and S. V. Rajashekararadhya, "Classification of Landsat 8 Imagery Using Kohonen ' s Self Organizing Maps and Learning Vector Quantization," in *Advances in Communication, Signal Processing, VLSI, and Embedded Systems*, no. January, 2020, pp. 445–462.

[16] F. M. Riese and S. Keller, "Introducing A Framework Of Self-Organizing Maps For Regression Of Soil Moisture With Hyperspectral Data," pp. 6151–6154, 2018.

[17] S. Keller, F. M. Riese, J. Stötzer, P. M. Maier, and S. Hinz, "Developing A Machine Learning Framework For Estimating Soil Moisture With VNIR Hyperspectral Data," *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.*, vol. 4, no. 1, pp. 101–108, 2018, doi: 10.5194/isprs-annals-IV-1-101-2018.

[18] P. Mallick, O. Ghosh, P. Seth, and A. Ghosh, "Kohonen's Self-organizing Map Optimizing Prediction of Gene Dependency for Cancer Mediating Biomarkers Partho," *Emerg. Technol. Data Min. Inf. Secur.*, pp. 863–870, 2019, doi: 10.1007/978-981-13-1501-5.

[19] M. J. Friedel, S. R. Wilson, M. E. Close, M. Buscema, P. Abraham, and L. Banasiak, "Comparison of four learning-based methods for predicting groundwater redox status," *J. Hydrol.*, vol. 580, no. September 2019, p. 124200, 2020, doi: 10.1016/j.jhydrol.2019.124200.

[20] M. Lourenco Baptista, E. M. P. Henriques, and K. Goebel, "A Self-Organizing Map and a Normalizing Multi-Layer Perceptron Approach to Baselining in Prognostics under Dynamic Regimes," *Neurocomputing*, vol. 456, pp. 268–287, 2021, doi: 10.1016/j.neucom.2021.05.031.

[21] W. Wang and Y. Lu, "Analysis of the Mean Absolute Error (MAE) and the Root Mean Square Error (RMSE) in Assessing Rounding Model," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 324, no. 1, 2018, doi: 10.1088/1757-899X/324/1/012049.

[22] C. Li *et al.*, "Estimating apple tree canopy chlorophyll content based on Sentinel-2A remote sensing imaging," *Sci. Rep.*, vol. 8, no. 1, pp. 1–10, 2018, doi: 10.1038/s41598-018-21963-0.

[23] M. H. Gholizadeh and A. M. Melesse, "Study on Spatiotemporal Variability of Water Quality Parameters in Florida Bay Using Remote Sensing," *J. Remote Sens. GIS*, vol. 06, no. 03, 2017, doi: 10.4172/2469-4134.1000207.

[24] F. M. Riese, S. Keller, and S. Hinz, "Supervised and semi-supervised self-organizing maps for regression and classification focusing on hyperspectral data," *Remote Sens.*, vol. 12, no. 1, 2020, doi: 10.3390/RS12010007.

[25] N. Verbeeck, R. M. Caprioli, and R. Van de Plas, "Unsupervised machine learning for exploratory data analysis in imaging mass spectrometry," *Mass Spectrom. Rev.*, vol. 39, no. 3, pp. 245–291, 2020, doi: 10.1002/mas.21602.

[26] J. Lever, M. Krzywinski, and N. Altman, "Points of Significance: Principal component analysis," *Nat. Methods*, vol. 14, no. 7, pp. 641–642, 2017, doi: 10.1038/nmeth.4346.

[27] D. Granato, J. S. Santos, G. B. Escher, B. L. Ferreira, and R. M. Maggio, "Use of principal component analysis (PCA) and hierarchical cluster analysis (HCA) for multivariate association between bioactive compounds and functional properties in foods: A critical perspective," *Trends Food Sci. Technol.*, vol. 72, no. 2018, pp. 83–90, 2018, doi: 10.1016/j.tifs.2017.12.006.

[28] A. Herrera, D. Ballabio, N. Navas, R. Todeschini, and C. Cardell, "Principal Component Analysis to interpret changes in chromatic parameters on paint dosimeters exposed long-term to urban air," *Chemom. Intell. Lab. Syst.*, vol. 167, pp. 113–122, 2017, doi: 10.1016/j.chemolab.2017.05.007.

[29] H. Shin, H. Jeong, J. Park, S. Hong, and Y. Choi, "Correlation between Cancerous Exosomes and Protein Markers Based on Surface-Enhanced Raman Spectroscopy (SERS) and Principal Component Analysis (PCA)," *ACS Sensors*, vol. 3, no. 12, pp. 2637–2643, 2018, doi: 10.1021/acssensors.8b01047.

[30] L. Wang, S. Wang, Z. Yuan, and L. Peng, "Analyzing potential tourist behavior using PCA and modified affinity propagation clustering based on Baidu index: Taking Beijing City as an example," *Data Sci. Manag.*, vol. 2, no. May, pp. 12–19, 2021, doi: 10.1016/j.dsm.2021.05.001.

[31] S. Asante-Okyere, C. Shen, Y. Y. Ziggah, M. M. Rulegeya, and X. Zhu, "Principal Component Analysis (PCA) Based Hybrid Models for the Accurate Estimation of Reservoir Water Saturation," *Comput. Geosci.*, p. 104555, 2020, doi: 10.1016/j.cageo.2020.104555.

[32] T. Bouwmans, S. Javed, H. Zhang, Z. Lin, and R. Otazo, "On the Applications of Robust PCA in Image and Video Processing," *Proc. IEEE*, vol. 106, no. 8, pp. 1427–1457, 2018, doi: 10.1109/JPROC.2018.2853589.

[33] N. Subba Rao, B. Sunitha, N. Adimalla, and M. Chaudhary, "Quality criteria for groundwater use from a rural part of Wanaparthy District, Telangana State, India, through ionic spatial distribution (ISD), entropy water quality index (EWQI) and principal component analysis (PCA)," *Environ. Geochem. Health*, vol. 42, no. 2, pp. 579–599, 2020, doi: 10.1007/s10653-019-00393-5.

[34] A. C. Eckert-Gallup, C. J. Sallaberry, A. R. Dallman, and V. S. Neary, "Application of principal component analysis (PCA) and improved joint probability distributions to the inverse first-order reliability method (I-FORM) for predicting extreme sea states," *Ocean Eng.*, vol. 112, pp. 307–319, 2016, doi: 10.1016/j.oceaneng.2015.12.018.

[35] T. Gergely, O. Georgiana, G. Pascal, F. Matei, and T. Salagean, "Statistical Analysis of a Digital Elevation Model Using Arcgis," *J. Young Sci.*, vol. IV, pp. 1–4, 2016.

[36] B. Balázs, T. Bíró, G. Dyke, S. K. Singh, and S. Szabó, "Extracting water-related features using reflectance data and principal component analysis of Landsat images," *Hydrol. Sci. J.*, vol. 0, no. 0, 2018, doi: 10.1080/02626667.2018.1425802.

[37] E. Sharaf and E. Din, "Enhancing the accuracy of retrieving quantities of turbidity and total suspended solids using Landsat-8-based-principal component analysis technique," *J. Spat. Sci.*, vol. 00, no. 00, pp. 1–20, 2019, doi: 10.1080/14498596.2019.1674197.

[38] L. Wang, "Research on Distributed Parallel Dimensionality Reduction Algorithm Based on PCA Algorithm," no. Itnec, pp. 1363–1367, 2019.

[39] Joint Research Center, "NDWI : Normalized Difference Water Index," 2011.

[40] H. Xu, "Modification of normalised difference water index (NDWI) to enhance open water features in remotely sensed imagery," *Int. J. Remote Sens.*, vol. 27, no. 14, pp. 3025–3033, 2006, doi: 10.1080/01431160600589179.

[41] D. Jiang, Q. Wang, F. Ding, J. Fu, and M. Hao, "Potential marginal land resources of cassava worldwide: A data-driven analysis," *Renew. Sustain. Energy Rev.*, vol. 104, no. December 2018, pp. 167–173, 2019, doi: 10.1016/j.rser.2019.01.024.

[42] A. Bronshtein, "Train/Test Split and Cross Validation in Python," 2017. https://towardsdatascience.com/train-test-split-and-cross-validation-in-python-80b61beca4b6.