Spatial model for predicting sugarcane crop productivity using support vector regression

A W M Gaffar¹ and I S Sitanggang¹

¹ Department of Computer Science, Faculty of Natural Science and Mathematics, Bogor Agricultural University (IPB), Jl. Meranti Wing 20 Level V, Kampus IPB Dramaga, Bogor, Indonesia 16680, e-mail: imas.sitanggang@ipb.ac.id

Abstract. Sugarcane which is the main ingredient of sugar production has a very high production demand so it must have a balance with the level of productivity. Sugarcane productivity in Indonesia can only reach 5.8 tons/ha which is still relatively low. In recent years sugarcane farmers have felt the decline in sugarcane production due to climate change which has affected the success of sugarcane planting so that sugar production has not been in line with national needs. This study aims to build a prediction model for the productivity of sugarcane crops in each area of Java Island based on climatic factors. The climate data used are collected from Indonesian Agency for Meteorology, Climatology, and Geophysics from 2006 to 2015. The method used is Support Vector Regression (SVR) using a radial base kernel function with a generalized neighbourhood matrix as a spatial model. The prediction model has RMSE of 0.1203954 and correlation of 0.9459743. The results of visualization of sugarcane productivity mapping show that areas in East Java have a higher value of productivity increase than areas in other provinces. Based on the results obtained, the model used has a quite good performance in predicting the productivity of sugarcane.

Keywords: climate, prediction model, sugarcane, support vector regression

1. Introduction
Sugarcane (saccharum sp.) is one of the most important plantation products that is used as the main ingredient in sugar production so it have high economic value [1]. In 2013 there was an increase about 3% in sugar consumption due to high industrial demand, which initially the consumption was only 2.6 million tons increased to 2.7 tons. The high demand for sugar production should be balanced by increasing the productivity of sugarcane [2].

According to the Central Statistics Agency (2014), the productivity of sugarcane plantations in Indonesia reached 56% of the area of sugarcane plantations in Indonesia which reached 194.9 hectares owned by the company and 247.8 hectares of smallholder plantations. Although every year there is a development of sugarcane harvest area in Indonesia, it affects the increase in sugarcane production but the amount of sugarcane productivity in Indonesia can only reach 5.8 tons/ha which is still relatively low.

East Java, Central Java, Lampung, West Java and Yogyakarta have a significant production contribution of sugarcane plants at 98.9% of the total amount of sugar production. East Java has a contribution of 69.75 %, it is the highest position as the most significant contributor to sugar production in Indonesia [3].
Climate and weather are factors that influence the success of sugarcane planting. The last few years sugarcane farmers feel the decline in sugarcane production due to weather changes so that sugar production does not meet national needs. In 2014, sugarcane production was targeted at 2.8 million tons but only 89.9% was achieved [4].

A study on crop productivity predictions has been carried out by Rahul Tripathi (2014) which predict productivity and yield of rice in Odisha and India by processing data on wide of area, production and productivity using the Autoregressive Integrated Moving Average (ARIMA) model. The result is an analysis of productivity trends along with the results of forecasting and validation [5]. Another study by Reza Septiawan (2012) developed a predictive model for the productivity of red chili and cayenne pepper in West Java and East Java based on climate data derived from weather stations. The prediction uses the Self-Organizing (SOM) clustering method from Kohonen [6]. The prediction of plant productivity in both studies still uses nonspatial data. The spatially related study was carried out by A. Podznoukhov (2011) to build spatial data on estimated avalanches in the Scottish region based on historical data, meteorological conditions and snowpack observations in the field using Support Vector Machine [7].

This study aims to develop a spatial model to predict the productivity of sugarcane crops in each area in Java Island based on climate influence using Support Vector Regression (SVR) technique. This study processes the spatial data and the results of sugarcane plantations with climate data in each area. The output of prediction model is the productivity of sugarcane crops with the aim to monitor the development of sugarcane productivity. The model can provide early warning to sugarcane farmers in case of decreasing productivity which is predicted to reduce the risk due to climate change.

2. Material and method

2.1. Data

The data used in this research are productivity, area, and yield of sugarcane crops in each area in the island of Java. Those annually data are collected from the website of the data center and Indonesian Plantation Statistics report (Sugarcane) published by the Ministry of Agriculture in 2006 to 2015. Climate data in the form of daily minimum and maximum temperature, humidity, rainfall and duration of sunshine are taken from Indonesian Agency for Meteorology, Climatology, and Geophysics at several UPT weather stations spread across Java. The attributes of data can be seen in Table 1.

2.2. Spatial prediction model using Support Vector Regression

Spatial data mining is the extraction of specific spatial information from spatial databases. Spatial data focus and refer to a dataset that has location or space aspects. Spatial data mining is usually used to detect image databases, remote sensing, navigation and route controllers, geographic information systems and various other sectors that require spatial information.

The characteristics of spatial data include the data has a very large amount of information so that an efficient algorithm is needed, have attributes that indicate spatial data such as latitude and longitude, there is ambiguity in spatial relationships or uncertainty of spatial location, and have a relationship with neighboring objects so that the data are more complicated [8].

Support Vector Machine is one of the methods developed in machine learning for classification and regression problems. Regression-based SVM (SVR) is a type of optimization model and can also be used to model nonlinear processes, thus the model can minimize error predictions and complexity model [9]. This method has an output in the form of continuous or real numbers so that it can overcome overfitting and performance that is better than the SVM algorithm. The SVM algorithm in the case of regression aims to build a hyperplane that is close to as much data as possible. Therefore, the optimal hyperplane is the one that matches all input data with the smallest possible error by mapping the input
vector into a higher dimension [10]. In this study, SVR is applied to predict the productivity of plantation crops, especially sugarcane.

Spatial prediction is carried out using a generalized neighborhood matrix w model. Regression coefficients are constant in some of the observations with appropriate parameter estimates. The neighborhood matrix is related to the dependence on the number of neighbors. The productivity prediction of sugarcane crops was built by prediction models using SVR algorithm by considering the most optimal hyperplane search [11]. The steps of Regression Vector Support algorithm [10]:

2.2.1. An optimal hyperplane. Function of the regression line as optimal hyperplane:

\[ f(x) = wx + b \quad w \in X, b \in R \]  

2.2.2. Quadratic Programming. Looking for the best hyperplane using Quadratic Programming optimization method by minimizing:

\[ \frac{1}{2} w^T w \]  

So, it will produce optimized functions:

\[ \min \left\{ \frac{1}{2} ||w||^2 \right\} \]  

The optimized function will make the function as small as possible so that it can control the function capacity. Assumed that all points are in the range \( f(x) \pm \varepsilon (\text{feasible}) \), in terms of infesibility, where there are some points that may come out of the range \( f(x) \pm \varepsilon (\text{feasible}) \) the slack variable \( \xi_i \) is added to overcome the not feasible, so that optimization is carried out formulated as follows:

\[ \min \left\{ \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{l} (\varepsilon_i + \xi_i^*) \right\} \]  

2.2.3. Lagrangian function. Build the following multiplier Lagrangian function:

\[ L = \frac{1}{2} \| w \|^2 - C \sum_{i=1}^{l} (\varepsilon_i + \xi_i^*) - \sum_{i=1}^{l} \lambda_i (\varepsilon_i + \xi_i^* - y_i + wx_i + b) - \sum_{i=1}^{l} \lambda_i^* (y_i + \varepsilon_i + \xi_i^* - y_i + w\xi_i + b) - \sum_{i=1}^{l} (\eta_i \xi_i + \eta_i^* \xi_i^*) \quad \text{and} \quad \lambda_i, \lambda_i^*, \eta_i, \eta_i^* \geq 0 \]  

2.2.4. Substitution of derivative calculations. To get the optimal solution, substitution calculation is substituted, as for the optimal w equation:

\[ w^* = \sum_{i=1}^{l} (\lambda_i - \lambda_i^*) x_i \]  

Then the optimal hyperplane is written below:

\[ f(x) = \sum_{i=1}^{l} (\lambda_i - \lambda_i^*) x_i + b^* \]  

2.2.5. Kernel function. The learning process in finding vector support points requires kernel function in solving the formulation of feature transformation into new spaces that are of a higher dimension efficiently [12].

3. Result and discussion

3.1. Pre-processing data

Preprocessing data begins by identifying the missing values in the dataset, using the summary (dataset) function. For climate datasets, since the weather station is not in all areas, the climate data in each area are determined based on the location of the closest weather station. The climate dataset which is the daily data must be averaged annually. The year data is ignored because in this study temporal prediction models were not carried out.

The spatial model in this study is shown by the neighbor matrix, where the productivity value of a district is influenced by the productivity value of the neighboring districts. For example in the Figure 1,
Bojonegoro has neighbors, namely Tuban, Blora, Ngawi, Madiun, Nganjuk, Jombang and Lamongan. Then the productivity value is normalized so that it has a value in the range from 0 to 1.

3.2. Build model prediction using Support Vector Regression
The dataset is divided into two, namely training data and testing data. The spatial prediction model is built using training data which is 80% of the entire dataset. Model building was done using the Support Vector Regression (SVR) technique with the radial base kernel function (RBF) with a default gamma of 0.125 which is processed using R.

The model that has been built was then tested using data testing to produce a predictive value of productivity. Then the comparison of the actual productivity value of the dataset and the value of productivity predictions was measured by the RMSE. The resulting RMSE is 0.1335017, and the predictive model correlation is 0.9371245. The graph in Figure 2 shows the comparison of the actual value symbolized by the red line and the prediction value symbolized by the blue line.

3.3. Evaluating prediction model
The model determination is a procedure to measure the average error in predicting the model using Root Mean Squared Error (RMSE) [13] which is formulated as follows:

\[ RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2} \]  

(8)

The multiple correlation coefficient \((R^2)\) is a relative measure of the relationship between one variable and another variable, so it can be known how strong the relationship between variables X and variable Y. With the formulation as follows [14] :

\[ R^2 = 1 - \frac{\sum (\hat{y} - y)^2}{\sum (y - \bar{y})^2} \]  

(9)

After building a spatial prediction model with the default gamma parameter of 0.125, then tuning parameters is performed to see the potential for performance improvement on the model. Tuning parameters are presented on gamma parameters with a range of values between 0.100 and 10 so that several models are formed which are then evaluated using the RMSE and the correlation coefficient which can be seen in Table 2.

3.4. Visualization predicting model
Visualization is a method to present data or problem into a graphic or image format so that it is easier to understand. The model that has been built previously using SVR produces prediction of the productivity of sugarcane crops in each area in Java Island. The results of the prediction are then visualized in the form of mapping in Java Island using QuantumGIS, wherein each area is symbolized by a color that means the range of values of sugarcane productivity in the area.

In the mapping as shown on Figure 3, four-color symbols have a range of values for the productivity of sugarcane. The exact white color in the area has no productivity value or does not have the sugarcane planting area. Yellow color indicates productivity value below 3000 Kg/Ha, light green has a range between 3000 to 4000 Kg/Ha, dark green means a range of 4000 to 5000 Kg/Ha. Whereas the blue color has a range between 5000 to 7000 Kg/Ha and the dark blue color has a productivity value above 7000 Kg/Ha.

4. Conclusion
This study results a spatial model to predict the productivity of sugarcane with a quite good performance, where the prediction results depend on neighboring matrix. The prediction model uses SVR techniques that process data using a radial base kernel function (RBF). RMSE of the model generated with the default gamma is 0.1335017, and the correlation is 0.9371245. Then tuning parameter was performed on gamma parameter value. The result is the model with lower RMSE, but its correlation is getting
higher. Based on the experiment, the best gamma value is 0.100 with the RMSE of 0.1203954 and the correlation coefficient of 0.9459743. Based on the results visualization, this study concludes that the sugarcane productivity in area in East Java has a higher value than those in other area in the province of Java.

5. Appendix

Table 1 Attribute of the data.

| Attribute Name    | Type    | Explanation                                                                 |
|-------------------|---------|------------------------------------------------------------------------------|
| Year              | Numeric | Year of calculating the productivity of sugarcane                           |
| Area              | Character | Area where sugarcane is produced                        |
| Province          | Character | Province of Area                                                               |
| Minimum temperature | Numeric | The minimum temperature in area is expressed in centigrade (%)              |
| Maximum temperature | Numeric | The maximum temperature in area is expressed in centigrade (%)              |
| Average humidity  | Numeric | The average humidity in an area                                              |
| Rainfall          | Numeric | Rainfall in an area stated in mm                                             |
| Duration of sunshine | Numeric | The duration of sunshine in an area within hours                           |
| Wide area         | Numeric | Area planted with sugarcane stated in Ha                                    |
| Production        | Numeric | Production of sugarcane which is expressed in tons                           |
| Productivity      | Numeric | The productivity value of sugarcane plants which is stated in Kg/Ha          |

Table 2 The results of tuning parameter.

| Parameter (Gamma) | RMSE       | Correlation |
|-------------------|------------|-------------|
| 0.100             | 0.1203954  | 0.9459743   |
| 0.125             | 0.1335017  | 0.9371245   |
| 0.50              | 0.1749498  | 0.7997868   |
| 1                 | 0.2170132  | 0.7034049   |
| 2                 | 0.3119699  | 0.5895436   |
| 5                 | 0.3708898  | 0.5001312   |
| 10                | 0.3929847  | 0.4671223   |

Figure 1. Example of determining matrix neighbors.
Figure 2. Graph comparison of actual values and predictive values.

Figure 3. Visualization of sugarcane productivity in Java.

References
[1] Mariska I and Rahayu S 2011 J Litbang Pertanian *Pengadaan bibit tebu melalui kultur jaringan*
[2] Diana N E 2017 J Ilmu Pertanian Indonesia *Pertumbuhan, Produktivitas, dan Rendemen Pertanaman Tebu Pertama (Plant Cane)* Pada Berbagai Paket Pemupukan
[3] Apriawan D C, Irham I and Mulyo J H 2014 Agro Ekonomi *Analisis Produksi Tebu dan Gula di PT. Perkebunan Nusantara VII (Persero)*
[4] Mustadjab M M, Fahriyah F and Hana I 2012 Agric Socio-Economics *Dampak Perubahan Iklim Terhadap Pendapatan Petani Tebu (Studi Kasus Pada Kptr (Koperasi Petani Tebu Rakyat)“Tani Mulya” Di Kabupaten Sidoarjo, Jawa Timur)*
[5] Tripathi R, Nayak A K, Raja R, Shahid M, Kumar A, Mohanty S, Panda B, Lal B and Gautan P 2014 Advances in Agriculture *Forecasting rice productivity and production of Odisha, India, using autoregressive integrated moving average model*
[6] Septiawan R, Komaruddin A, Sulistyia B, Alfi N and Shanmuganathan S 2012 Prediction model for chilli productivity based on climate and productivity data (Bandung: Geise Parahyangan Catholic University)
[7] Pozdnoukhov A, Matasci G, Kanevski M and Purves R S 2011 Nat Hazards Earth Syst Sci. *Spatio-temporal avalanche forecasting with support vector machines*
[8] Bist A and Faridi M A 2017 Int. J. of Eng. Trends and Tech. Survey: On Spatial Data Mining 46 327
[9] Sánchez A S, Nieto G, Fernández R, Díaz D and Rodríguez I 2011 Math Comput Model Application of an SVM-based regression model to the air quality study at local scale in the Avilés urban area (Spain) 54 5–6
[10] Trafalis T B and Ince H 2000 Support vector machine for regression and applications to financial forecasting (Italy:IEEE) 6
[11] Cheng T, Wang J and Li X 2007 In: Proc Geocomputation The support vector machine for nonlinear spatio-temporal regression
[12] Maji S, Berg A C and Malik J 2013 IEEE Trans Pattern Anal Mach Intell. Efficient classification for additive kernel SVMs 35 1
[13] Esmaeili M, Salimi A, Drebenstedt C, Abbaszadeh M and Bazzazi A Arab J Geosci Application of PCA, SVR, and ANFIS for modeling of rock fragmentation 8 9
[14] Draper N R and Smith H 2014 Applied regression analysis (Canada:John Wiley & Sons) 326