International Journal of Computer Applications Technology and Research
Volume 5– Issue 5, 253 - 259, 2016, ISSN:- 2319–8656

Forest Area Estimation in Kutai Nasional Park of East Kalimantan Using Computer System Aplication based Genetic Algorithm-Support Vector Machine

Lapu Tombilayuk
Informatics Department, Sekolah Tinggi Teknologi Bontang,
Bontang East Kalimantan, Indonesia

Abstract : The paper presents design of computer system application for the forest area estimation using the combination of genetic algorithm (GA) and support vector machine SVM) methods in Kutai National park of East Borneo. The considering variables such as reboization concerning natural green, forest fire, encroachment and illegal logging activities are the basic data for our proposed design. In the development of design computer system application, the Unified Modeling Language is adopted with the use case, activity diagrams and sequence diagrams are systematically followed in order to keep the purpose of design on track. The supporting instrument for this research is the language programming of Borland Delphi 7 and MySQL database. The accuracy of area estimation result is compared with the actual data using mean absolute percentage error (MAPE).

Keywords: GA-SVM methods, UML, Borland Delphi, MySQL database, white box testing.

1. Introduction

Indonesia has continuously experienced in deforestation area where the majority of forest damaging occurs in Borneo and Sumatera. The causes of deforestation are forest mismanagement, illegal logging, forest fire and forest opening for farming and mining purposes. The effects of forest damaging can be locally, such as ecologic disaster, flood, soil avalanche and can be globally as well, such as drought and other global warming effects. The government is actually proactive to do the forest conservation program with policies, for instance the protection of primary natural and peat forests and the strict regulation of forest utilization for mining purpose. The government regulation on forest conservation is highly protected by law; in fact, deforestation seems uncontrollable. The organization of forest watch Indonesia (FWI) reported that the forest in Java will be used up in 2020 with small width area remaining in Bali and Nusa Tenggara (0.08 million h.a), Sulawesi (7.2 million h.a), Sumatera (7.72 million h.a), Borneo (21.29 million h.a) and Papua (33.45 million h.a). The rate of deforestation is about 1.51 million h.a per year in the period of 2000-2009 with the highest rate of 0.55 million h.a per year occurs in Borneo.

Forest area estimation method has gained important awareness of researchers with variety methods. The statistic methods are still dominant in this topic. With the hierarchy of Bayesian model, the forest of area can be accurately classified up to 88%. Another approach by the forest sampling method is used to estimate the forest canopy cove according to probability theory. In addition, the forest cover estimation based the importance vegetation type has been proposed using K-means method on the time series data. The research is focused on the clustering of different type of vegetation. Meanwhile, the estimation of forest canopy by comparison of the field measurement technique has been conducted by pictures result of digital camera.

The results of the previous research motivate us to find the best method to estimate the forest area based on the conditions of reboization concerning natural green, forest fire, encroachment and illegal logging activities. The parametric method by mean mathematic model and statistic, such as autoregressive integrated moving average (ARIMA) is less suitable for this case study due to the difficulty in modeling irregular and variable number of data. On the other hand, the non-parametric method with exponential technique is only superior for the short-term forecasting and the results may not be confirmed optimal. In addition, the artificial neural network is facing difficulty to provide solution with time-series data.

2. Configuration of proposed system

The time series estimation method has continuously attracted serious attention from scientific community. The method is basically part of computational intelligence utilizing historical data to solve estimation problems. This kind of technique is possible supporting fromthe advanced computational technique and information technology. One of the computational techniques based machine learning is the Support Vector Machine (SVM). This method is superior to classify the input data set from the minimum to the maximum values. The classification results are used as chromosome for the genetic algorithm (GA) operation. This combination accelerates the computational efforts and provides high accuracy estimation compared to the only GA process. In this case, the output genetic algorithm (GA) is the optimal

estimated value of forest area. It also implies that the uncorrelated data classification is avoided, as results only the important inputs are considered by the implementation of GA-SVM method. In this research, the main configuration of the estimation method for the forest area in Kutai National park, East Borneo is divided into the input database of including positive, like reboization concerning natural green and negative causes, such as forest fire, encroachment and illegal logging activities that affecting to the forest area, processing database using GA-SVM methods and the area forest estimation output. The database system is stored using MySQL software system. The genetic algorithm utilizes these data to initialize parameter and to generate the population. Later, the support vector machine method evaluates the fitness function in order obtain the data set. Then, the data set is reselected to obtain two chromosomes with the best fitness function by the genetic algorithm. By this approach, the only correlated data is used for the GA process, results in high accuracy estimation. The process and evaluation continue with crossover and mutation of new generation until it convergences at the best intent of fitness function.

The input database is taken directly from Kutai National Park office during the last 10 years (between 2003 and 2012) about data record of reboization, forest fire, encroachment and illegal logging activities.

These data is quite random and irregular due to the cause combination between the nature and human activities. In these

data, the average forest damaging area based on the negative causes is about 125.68 h.a, with the mean area of reboization is about 54.5 h.a. In 2013, the total forest area of Kutai National ark, East Borneo is about 198,629 h.a with damaging area about 711.8 h.a. Based on this reason, it is important to provide some tools to estimate the forest area for some time in the future, so that it becomes reference to measure and to accelerate the green activities inside the National Park.

The forest area estimation is performed by the implementation of combination between the genetic algorithm (GA) and support vector machine (SVM) method based on the storing data in the database system. The superiority of this method is in the capability of SVM method to divide vector space into hyperplane according to the data trend from small area to wide area classes. The algorithm combination for such estimation technique is able to provide better solution compared to other estimation techniques, such as artificial neural network because the error and generalization of the SVM method is not depending on the input vector. The complete process of GA-SVM methods is shown by the pseudo-code as follows:

1) Initialization process by the genetic algorithm to select the chromosome candidates of data stored in database.
2) Initialization of data set by running the SVM method to search all data set about reboization, forest fire, encroachment and illegal logging activities between 2003 and 2012 in the data base system.
3) Run the polynomial Kernel function to classify the non-linear data into two classes by $(\boldsymbol{XT.Xi} + 1)P$. The results are about the influence causes, measured from the most to the less value impacts.
4) The training process of all selected data set by means the influenced parameters to the forest area. The selection is aimed to obtain the hyperplane of $y(x)f(x)=1$ that separates 2 classes. The candidates of support vector are:

$$y(x)f(x) \leq +\beta+1 \text{ and } y(x)f(x) \geq 1 \quad (1)$$

where $\mathbf{Xa=Xa \cup Xb}$ and β is the arbitrarily determined value by users.

If the re-training process is conducted, then the previous training results are improved by taking only some data ($X \varepsilon Xa$) as the support vector candidates.

5) The data classification after training process is divided into $xi.w+b \geq 1$ for the 1st class and $xi.w+b \leq -1$ for the 2nd class.
6) Chromosomes selection. The best chromosome is usually selected by objective function evaluation with defined high probability. The roulette-wheel method is used in this step.
7) Fitness function evaluation by:

$$Fitness = C - f(x) \text{ or } Fitness = \frac{C}{F(x)+ \epsilon} \quad (2)$$

where C is a constant and Є is small number to avoid zero division. In this research, The fitness function = forest fire + encroachment + illegal logging – rebozation

8) Selection process with Linier Fitness Ranking (LFR).
9) Cross-Over process. One-cut point method is used to exchange the gen of the parent chromosome with cross-over probability (Pc) of 0.25.

```
Begin
  k ← 0;
  While (k < populasi) do
  R[k] ← random [0 - 1];                (3)
  If (R[k] < Pc) then
  Select Chromosome[k] as parent;
End;
```

10) Mutation process. The mutation rate is specified at 0.1.
11) Population replacement by the new generation
12) If the new generation convergences to the optimal solution, the overall process stop; otherwise the process is repeated from no. 10.

The above pseudocode is translated into language programming of Borland Delphi 7. The mean absolute percentage error (MAPE) is used to validate the accuracy in output measurement. In addition, the white box testing is used to evaluate all computer logic programming by checking the logical iteration and to assess the overall data used in the simulation.

3. Design of Computer Application for Forest Area Estimation

In this section, it will be explained about the designs information related to the unified modeling language (UML), database and user interface. The detailed information of such design is provided as follows:

3.1 *Design of Unified Modeling Language (UML)*

The unified Modeling Language (UML) is the standar design and documentation of software system with the capability of software application modelling according to the hardware configuration, operating system and it has the flexibility ain network design and language programming. The UML diagram can be built-up using use case, class, activity, sequence and statechart diagrams. With the UML design, the programmers may expect the model runs perfectly and accurately with considering the scalability, robustness and security. The use case diagram represents the expected functionality of designed system and also can be considered as the interaction mechanism between the user and the system. It consists of case login to access the system, case view menu with several menu options, case input data of such the causes that influences to forest area, case analysis by means the GA-SVM method and the case output as the estimated forest area. The input data processing may be performed by an administrator and the output data is utilized by the National Park Authority in order to anticipate the negative causes and continuing promote the green activities. The use case diagram is shown in Figure 1.

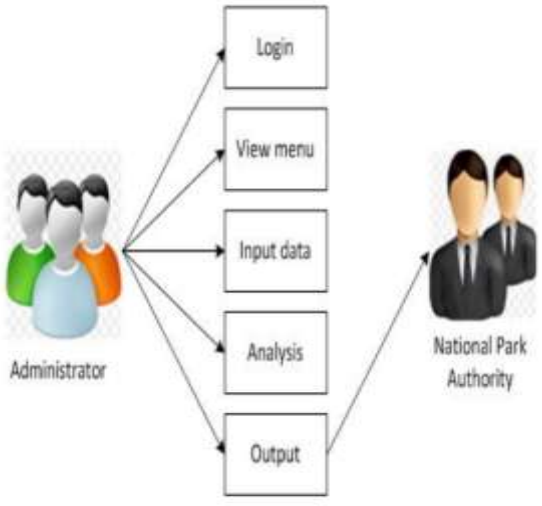

Figure 1. Use Case Diagram (4)

The class diagram represents the class of structure and its description, package and object including the connectivity within the application design. Figure 2 shows the class description of our proposed application. There are 3 data classes, i.e input data, analysis and report. The class of input data consists of causes identity, year, location, causes and width area fields. Meanwhile, the class of analysis is the fields of number, year, causes and width area. The last class is the class of report which comprises of fields of report identity and its type. The chart of class diagram is shown in Figure 2.

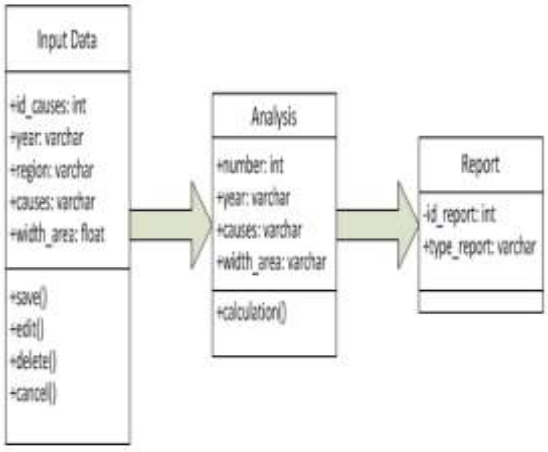

Figure 2. Class diagram

The next diagram in UML development is the activity diagram that describes the activity path inside the designed system, to how the system starts to end including the possibility decision that might be appear.
Amongst the developed diagram, the activity diagram is the special state diagram where most of the states are an action, triggered transition and internal processing by the previous states. In this research, the activity diagram explains the system process from the data inputting by an administrator, data processing using GA-SVM method and the estimation output utilization by the National park authority. The activity diagram is shown in Figure 3.

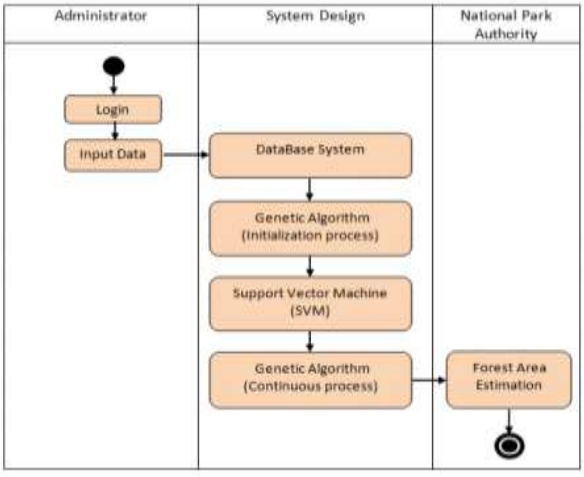

Figure 3. Actifity Diagram

The object interaction inside and surrounding the system including the user, display and so on is described by sequential diagram in the forms of messages per time value. The sequential diagram can also be used to express the scenarios or steps that should be followed according to the output events. It may start from the causes that trigger the next activity, to what kind of internal changes and typical. Each object including administrator has vertical life line. The message is depicted by arrows from an object to other objects. In the next phase design, such messages are mapped into operation or method of classes. Figure 4 explains the object behavior when the operator interacts with application system. Administrator needs login to input the causes that influence to the total forest area into the database. Later, this data is processed by the GA-SVM method to yield total forest area estimation which can be used by the National park authority.

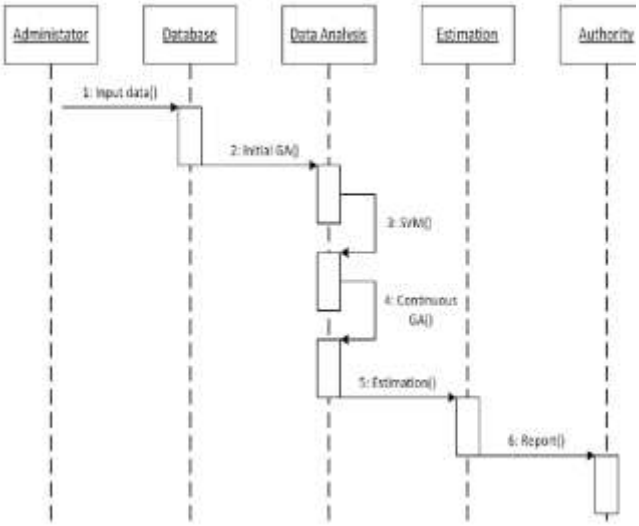

Figure 4. Sequential diagram

To support all diagrams process, it requires the state chart diagram input data that describes transition states or changes of object inside the systems. The state chart diagram is shown in Figure 5. In this figure, if the operators makes mistake during data inputting process, they have opportunity to do editing then resave the data into the database. If the cancelation of inputting data occurs, the system will terminate. The inputting data process into the database keeps going until the data is finished and saved. The remained process is just waiting the execution of GA-SVM method to yield the estimated forest area. In addition, the state chart diagram may describe the data analysis using GA-SVM method and report analysis as the outcome of algorithm.

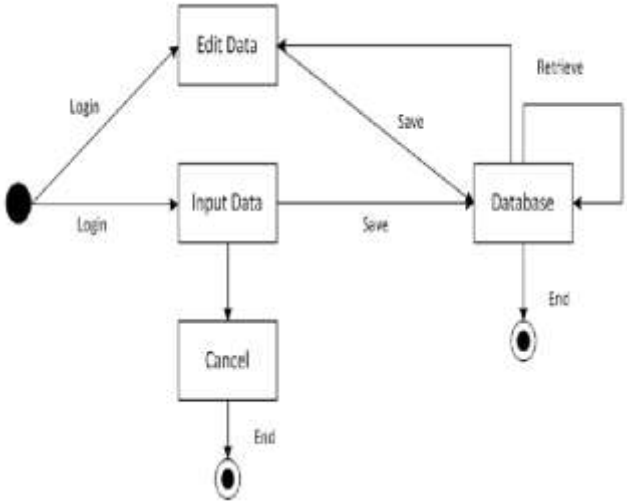

Figure 5. Statechart diagram of input and edit data

3.2 *Design of database*

The database design is used to map the conceptual model into the basis data model. In this research, the structure data file about the causes affecting the forest area located in the database is shown in Table 1. There are five field names but the most important field (primary) is the field of number. When transforming these data into terms of generation data in Genetic Algorithm called the data identity of generation, there are four field names. Again, the most important field (primary) is the field of number. The data identity of generation is shown in Table 2.

Table 1. Data identity of causes influenced to the forest area

Field name	Type	Length	Key	Declaration
Number	Int	11	Primary	The code of causes
Year	varchart	5	-	The year of events
Region	varchart	50	-	Forest region
Causes	varchart	50	-	Causes to the forest area
Width area	float	0	-	Damaged forest area

Table 2. Data identity of generation

Field name	Type	Length	Key	Declaration
Number	Int	11	Primary	The code of causes
Generation	varchart	5	-	Number of generation
Causes	varchart	50	-	Causes to the forest area
Width area	floa	0	-	Damaged forest area

In addition, all data in the database system in cooperating with software system is stored with MySQL basis management system. MySQL system is free software (open source) under the license of general public license (GPL) both the source code and executable program. The other advantage of MySQL are portability, multiuser, flexible tuning performance, high security and scalable. Therefore, design such computer application system is nowadays very convenient with maximal performance.

3.3. *Design of User interface*

The proposed user interface is the displays of the main menu, input data, initial data, generation and estimation. These displays are explained as follows. The display of main menu for the forest area estimation in Kutai National Park of East Borneo as shown in Figure 6 contains the communicative information related to input data, analysis and report. The display explains the main front of application on how the users interact to the system application simply. When the main display is active, the ser may input data by simple click. If they continue for the analysis and report, they may just precede the previous activity.

Figure 6. Display of main menu

In the inputting data process, the users may input data of year, region, and causes both positive and negative. The year data is specified from 2003 to 2012 with three definite regions where region I is called Suka Rahmat, region II is Sangatta and region III is Manamang. The last required data are the causes and their affected total area. The causes influenced to the forest area in the park are reboization, forest fire, encroachment and illegal logging activities. The page display for the input data process is shown in Figure 7.

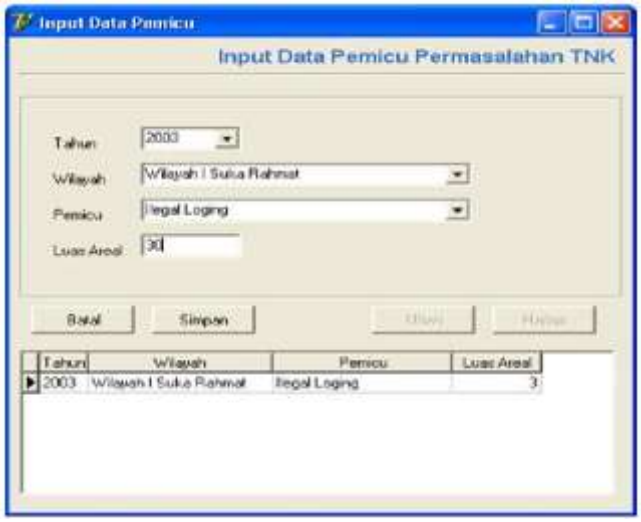

Figure 7. Interface display for input data

After the input data process is complete, the user may check their data stored in the database. The typical initial data stored in database system is shown in Figure 8. It is clearly shown that there are certain damaged forest area affected, forest fire, encroachment and illegal logging activities even though the reboization contributes positively in this matter. Total forest area in Kutai National Park, East Borneo is 198,629 h.a. However, it has been recently found that the total forest area reduced to 197,917 h.a by considering the above causes. In the current display, the users may continue the process until they obtain the estimated area in future years.

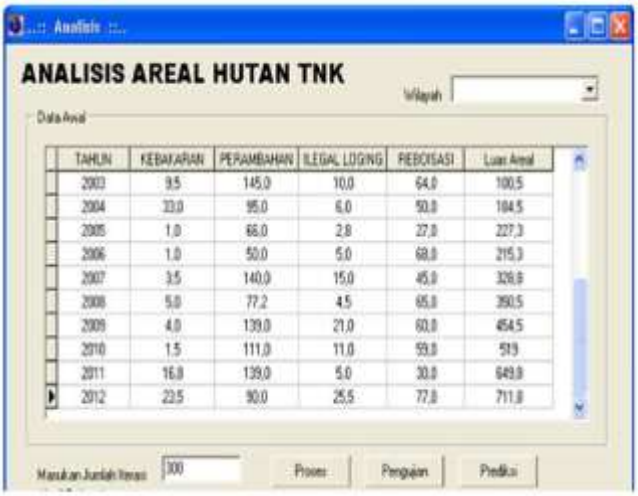

Figure 8. Initial input data display in database system

If the process continues, then the users may receive information about the estimation of forest area in the following years. In this simulation, the maximum width of forest area (the chromosome value) of the generation is the estimation output. It is due the consideration of the 10 years previous data to predict the forest area in the year after. For example in Figure 9, the chromosome value change from the initial value (data from 2003 to 2012) to other number specified that the damaged forest area in 2013 is 784.7 h.a. With the same consideration in the following generation, it will be the estimation results of the year 2014, and so on. With this application system, the estimated damaged forest area in the years of 2014, 2015 and 2016 are 905.1 h.a, 1,444.7 h.a and 2,211.2 h.a, respectively. These results are obtained with assumptions that there is no updating data from the initial data condition. The estimated results in 2016 may less than the above estimated number if the reboization activity increases and one of the negative causes can be pressed down.

Figure 9. Typical display of damage forest area estimation in 2013

If the process continues, then the users may receive information about the estimation of forest area in the following years. In this simulation, the maximum width of forest area (the chromosome value) of the generation is the estimation output. It is due the consideration of the 10 years previous data to predict the forest area in the year after. For example in Figure 10, the chromosome value change from the initial value (data from 2003 to 2012) to other number specified that the damaged forest area in 2013 is 784.7 h.a. With the same consideration in the following generation, it will be the estimation results of the year 2014, and so on. With this application system, the estimated damaged forest area in the years of 2014, 2015 and 2016 are 905.1 h.a, 1,444.7 h.a and 2,211.2 h.a, respectively. These results are obtained with assumptions that there is no updating data from the initial data condition. The estimated results in 2016 may less than the above estimated number if the reboization activity increases and one of the negative causes can be pressed down.

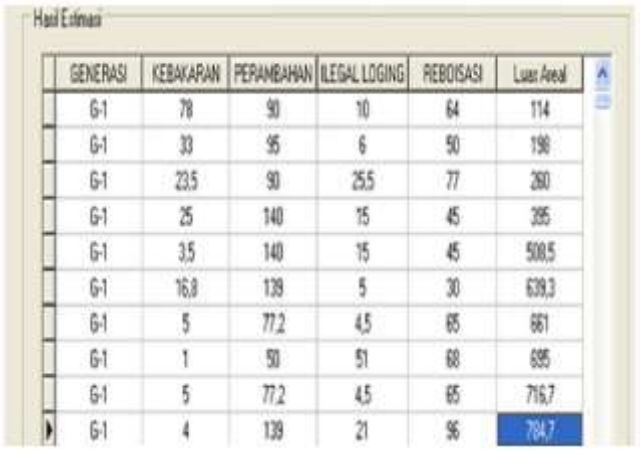

Figure 10. Typical display of damage forest area estimation in 2013

The design of application program is complete with the display of graph estimation as shown in Figure 11. In this figure, it is shown that in certain period the damage forest area decreases. For instance, there is significant reduction of forest area to about 939 h.a in 2017 and 358.4 h.a in 2020 from the previous year conditions. However, this condition is some kind of transition states because the total damaged area rises again to 1,883.9 h.a and 2,341.3 h.a in the years of 2018 and 2019, respectively. As previously mentioned that the current forest area of Kutai National Park, East Borneo is 198,629 h.a. If the prediction process continues without any changes in the input data by means there is no positive action for the green activity, the forest area is used up in 2061 because the estimated damage area in this year has reached 200,129 h.a which is far beyond the current forest area. Obtaining such number is important for the local people, park authority and global society as the reference to do positive action for the forest conservation since the Kutai National Park of East Borneo is 'the lung of world'.

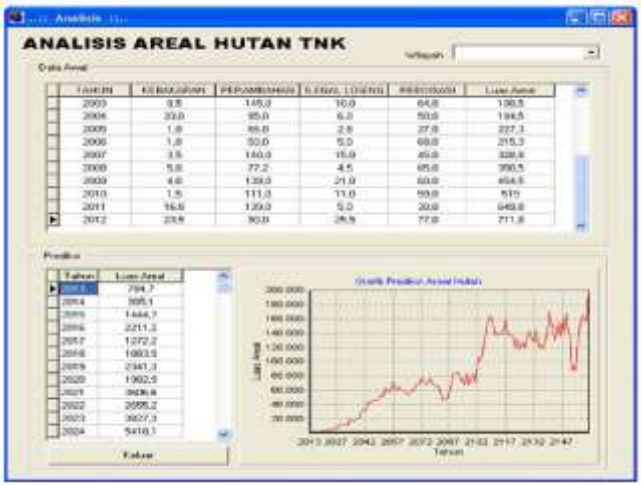

Figure 11. Typical display of estimation results with graph estimation

4. Simulation results and discussion

In the design of system application, the high accuracy estimation or prediction is the most important aspect to be considered in order to guarantee the outcomes are on the right value even the data input collection is abruptly changed. The accuracy assessment in this research is by measuring the Mean Absolute Percentage Error (MAPE) between the estimated area as the output of the designed system application and the actual area. In addition, the estimated output area is also compared with the conventional linear regression with similar MAPE performance index measurement.

The first scenario is the comparison between the estimated area and the actual area. For the easy comparison, the data of previous five years from 2003 to 2007 is arbitrarily selected as the initial input data (training data) because we have the figure actual data from 2008 to 2012 for the validation data. The result of simulation under scenario is shown in Fig. 11. In the simulation results, we have comparing data between estimated area and actual area from 2008 to 2012. These data is evaluated by Mean Absolute Percentage Error (MAPE) equation as follows:

$$\text{MAPE} = \frac{\sum_t^N \left| \frac{A_t - E_t}{A_t} \right|}{N} x100\% \quad (4)$$

where At is the actual area in the t year, Et is the estimate area in the t year and N is the period of data evaluation (N=5) in this research. The MAPE index performance calculation is summarized in Table III. For the five years period of estimation, the average of MAPE is about 2.1 % with the maximum difference of 102.8 h.a in 2008 and minimum difference of 8.8 h.a in 2011. It means that the proposed system application to estimate the forest area of Kutai National Park, East Borneo is guarantee small. Therefore, the estimated forest area until the year of 2020 may present proper results. Typical display of estimation results for the following years has been shown in Figure 12.

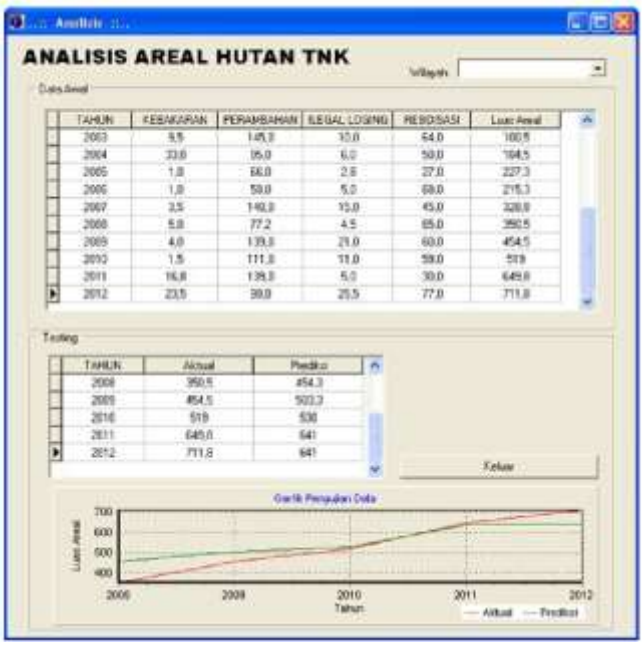

Figure 12. Simulation result for the estimated and actual area comparison

Table 3. MAPE performance index measurement

Year	Estimated area (h.a)	Actual area (h.a)	Difference (h.a)	MAPE (%)
2008	453.3	350.5	102.8	5.9
2009	503.3	454,5	48.8	2.1
2010	530.0	519.0	11.0	0.4
2011	641.0	649.8	8.8	0.3
2012	641,0	711.8	70.8	2.0

The last performance testing to our proposed system design is the White Box Testing. The test is conducted to see the module contents in order to evaluate the property of coding program. The test may cover the logical statements and their decision, iteration process within its constraints and the overall internal data structure to guarantee the validity of program. Basically, the white box testing is the path testing to allow the programmers to measure the logical complexity of procedural design and to use this measurement as the guidance to define the path set. The white box testing of this research is shown in Figure 13. There are 5 regions, denoted with R1= initial population, R2= chromosome selection, R3= crossover, R4= mutation and R5 = generation iteration. According to *the* simulator Flowgraph application, the Edge number is 14 and the Node number is 11; therefore the Cyclometric Complexity (CC) is equal to 5. Also, we obtain 5 paths from Fig. 12. which are Path 1: A – B – C – D – E – D – E – F – G – H – I – J – K; Path 2:A – B – C – D – E – F – G – F – G – H – I – J – K; Path 3: A – B – C – D – E – F – G – H – I – H – I – J – K; Path 4: A – B – C – D – E – F – G – H – I – J – C – D – E – F –G – H – I – J – K and Path 5: A – B – C – D – E – F – G – H – I – K. Because of the proposed system application has 5 regions, 5 CCs and 5 paths, the analysis application can be claimed to be properly correct.

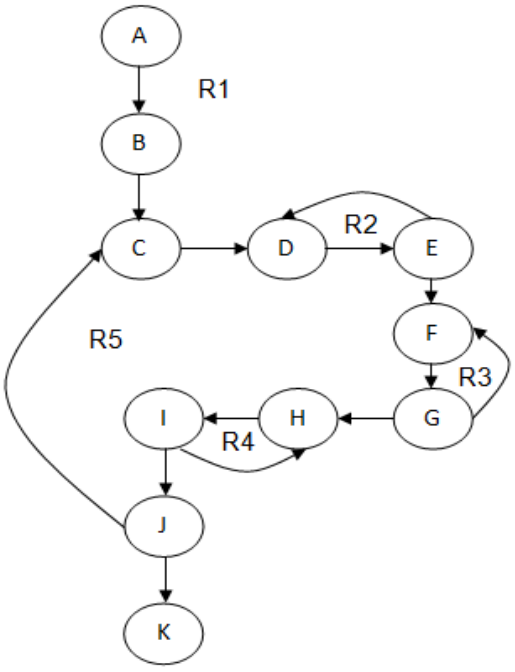

Figure 13. The flow graph in white box testing

5. Conclusion

The paper presents the design computer system application to estimate the forest area in Kutai National Park, East Borneo based the combination of genetic algorithm (GA) and support vector machine (SVM). In the development of design system application, the Unified Modeling Language is adopted with the case, activity and sequence diagrams are systematically followed in order to keep the purpose of design on track considering variables such as reboization concerning natural green, forest fire, encroachment and illegal logging as the input data in database system. The supporting instrument for this research is the language programming of Borland Delphi 7 and MySQL database. Without any actions to the initial data by means there is no significant effort to do the forest conservation program by the park authority, the forest area will be finished by the year of 2061. It is very dangerous situation to the world ecosystem because this park is 'the lung of the world'. Provision initial data information is very urgently to save our environment. The accuracy of area estimation result is compared with the actual data using mean absolute percentage error (MAPE) with the average error of 2.1%. In addition, the white box testing to calculate region, cyclometric complexity and path is implemented to confirm the correctness of the logic algorithm of design application program.

References

[1] Wirenro Sumargo, Soelthon Gussedya Nanggara, Frionny Nainggolan, Isnenti Apriani: "Potret Keadaan Hutan Indonesia Periode tahun 2000-2009", Forest Watch Indonesia, 2011, pp. 20-22.

[2] Peraturan Pemerintah Republik Indonesia Nomor 60 Tahun 2012: "Perubahan Atas Peraturan Pemerintah Nomor 10 Tahun 2010 Tentang Tata Cara Perubahan Peruntukan dan Fungsi Kawasan Hutan", 2012.

[3] Dudy Subagdja: "Nasib Hutan Kita dan Kebijakan Ekonomi Hijauh", Berita Kompasiana, 29 Maret 2013.

[4] Christina Basaria S.: "Kajian Kelestarian Tegakan Dan Produksi Kayu Jati Jangka Panjang KPH Bojonegoro Perum Perhutani Unit II Jawa Timur", Institut Pertanian Bogor, 2009.

[5] Dyah Pratiwi, et.al: "Penghitungan laju luas area hutan berbasis algoritma segmentasi warna local", Konferensi Nasional Sistem Informasi 2013, pp. 471-475.

[6] F. Deppe: "Forest Area Estimation Using Sample Survey and Landsat MSS and TM Data", Photogrammetric Engineering & Remote Sensing, Vol.6+, No.4, 1998, pp. 285-292.

[7] Dang Khoi dan Yuji Murayama: "Forecasting Areas Vulnerable to Forest Conversion in the Tam Dao National Park Region", Remote Sens. Vol. 2, 2010, pp. 1249-1272.

[8] Loghman Ghahramany, Pariz Fatehi, Hedayat Ghazanfari: "Estimation of Basal Area in West Oak Forests of Iran Using Remote Sensing Imagery", International Journal of Geosciences, Vol. 3, 2012, pp. 398-403.

[9] Oliver Diedershagen, Barbara Koch: "Automatic Estimation Of Forest Inventory Parameters Based on Lidar, Multi-Spectral and Fogis Data", Holgen Weinacker, University Freiburg, Germany, 2003, pp. 4-13.

[10] Andrew O. Finley, et.al.: "A Bayesian approach to multi-source forest area estimation", USA, Environ Ecol Stat, Vol. 15, 2008, pp.241–258.

[11] Farshad Keivan Behjou, Mahbobeh Foshat: "Using A Sampling Method for Estimation of Forest Canopy cover", International Journal of Agriculture: Research and Review. Vol., 3 (2), 2013, pp. 217-222.

[12] Anuj Karpatne, et.al.:, "Importance of Vegetation Type in Forest Cover Estimation", University of Minnesota, 2010.

[13] Lauri Korhonen, Kari T. Khorhonen, Miina Rautianen and Pauline Stenberg: "Estimation of Forest Canopy Cover: a Comparison of field measurement Techniques", Silva Fennica, Vol. 40, No. 4, 2006, pp.577-588.

[14] Thi Nguyen, Lee Gordon-Brown, Peter Wheeler, Jim Peterson: "GA-SVM Based Framework for Time Series Forecasting", the fifth International Conference on Natural Computation 2009, pp.493-498.