Stepwise regression of agarwood oil significant chemical compounds into four quality differentiation

Siti Mariatul Hazwa Mohd Huzir¹, Aqib Fawwaz Mohd Amidon², Anis Hazirah ‘Izzati Hasnu Al-Hadi², Nurlaila Ismail¹, Zakiah Mohd Yusoff¹, Saiful Nizam Tajuddin¹, Mohd Nasir Taib³,⁴

¹Department of System, School of Electrical Engineering, College of Engineering, Universiti Teknologi MARA, Pasir Gudang, Malaysia
²Department of System, School of Electrical Engineering, College of Engineering, Universiti Teknologi MARA, Shah Alam, Malaysia
³Bioaromatic Research Centre of Excellence (BARCE), Universiti Malaysia Pahang (UMP), Gambang, Pahang, Malaysia
⁴Malaysia Institute of Transport (MITRANS), Universiti Teknologi MARA, Shah Alam, Malaysia

Article Info

Article history:
Received Aug 3, 2022
Revised Sep 20, 2022
Accepted Oct 31, 2022

Keywords:
Agarwood oil
Significant compounds
Stepwise regression
Grade classification
Intelligent model

ABSTRACT

This paper gives precise summary on the application of stepwise regression model based upon the pre-process analysis of boxplot for four chemical compounds into four different qualities of agarwood oil. In the global market, agarwood oil is acknowledged as a pricey and valuable nature product owing to its benefits. Unfortunately, there is no standard grading method for agarwood oil grade classification. Intelligent model in grading the quality of agarwood oil is crucial as one of the efforts to classify the agarwood quality. The main model chosen in this study is stepwise regression by concerned specific parameter which is the value of correlation coefficient, R². To achieve this goal, four out of eleven significant compounds of agarwood oil that consist of 660 data samples from low, medium low, medium high and high quality are representing the input. The independent variables are X1, X2, X3 and X4 which refer to the ɤ-Eudesmol, 10-epi-ɤ-Eudesmol, β-agarofuran and dihydrocollumellarin compounds, respectively. MATLAB software version r2015a has been chosen as the simulation platform for this research work. The result showed that the stepwise regression model has a correlation coefficient of 0.756 and p-value less than 0.05 significance level which successfully passed the performance criteria toward regression value.

This is an open access article under the CC BY-SA license.

1. INTRODUCTION

Agarwood oil is particularly valuable as natural products used as incense, fragrant, shampoo, traditional medicine (healing stomach complaints, diarrhoea, lungs and liver pain), and for perfumery (especially the dark colors) [1]-[5]. In fact, the infected heartwood of Aqualaria species is the most expensive oil in the market [3], [5], [6]. The use of agarwood oil for variety purposes has recently grown in popularity. Market statistics indicate strong growth in the purchase of agarwood oil in the Middle East countries (United Arab Emirates, Saudi Arabia), China and Japan. Along with this growth in consumer sales, the need for agarwood oil productivity has increased including in Malaysia [3], [7], [8]. To cope with that, the population of agarwood plantations in Malaysia is rapidly expanding in Perak, Terengganu, Kelantan, Pahang even around the country [9], [10].
The agarwood oils are traded and priced differently depending on its grade which ranges from low, medium low, medium high and high quality [11], [12]. Traditionally, agarwood oil has been graded simply on the basis of its colour, resin content, taste, long-lasting aroma, and density by a human sensory panel for a long time [10], [11]. However, the sensory evaluation method is a little off. There is no guarantee that grading essential oils depending on human sensory evaluation can guarantee its purity or quality. The human trained grader technique has a considerable disadvantage in terms of objectivity and consistency when working with multiple samples at once, resulting in a labor-intensive and time-consuming procedure [1], [11], [13]. Many innovative technologies for enhancing the stability and availability of essential oils have emerged to address these inadequacies of essential oil products [14], [15]. Traditional systems are gradually being superseded by modern systems of quality assessment grading with new indexes ranging from qualitative to quantitative analysis [10]. A scientific method is an alternative solution to handling the grading issue. With new development of data analysis, there is several platforms where agarwood oil quality classification can be done solely based on their chemical profiles using intelligent methods, enabling essential oils to be classified into their respective classes (low, medium low, medium high, or high quality) and the findings to be measured accurately. Several researchers have been proposed machine learning techniques to verify the quality of agarwood oil such as artificial neural network (ANN), linear regression, k-Nearest Neighbor (k-NN), selforganizing map (SOM) and OVO multiclass support vector machine (SVM) [1], [13], [16]-[18]. Studies on the agarwood oil seven chemical compounds (β-agarofuran, 10-epi-Y-eudesmol, Y-Eudesmol, Eudesmol, Hexadecanol, α-agarofuran and Longifolol) have reported value of correlation coefficient, R is equal to 1 at hidden neurons number 2 outperforms shows the best performance since the mean squared error (MSE) value is the lowest compared to other neurons with 7.69x10^15 [16]. The SOM model found three chemical compounds which are α-agarofuran, β-agarofuran and 10-epi-γ-eudesmol were determined to be significant compounds for agarwood oil [13]. Majority of previous research work only include two qualities which are low and high quality.

Hence, according to the intelligent techniques mentioned above, this research work more focuses on using stepwise regression instead of linear regression as the main model to classify the grade of agarwood oil into four qualities (low, medium low, medium high, and high quality), as suggested by paper [16]. The advantage on stepwise regression approach is that it is improves the efficacy of agarwood grading by sorting the significant chemical compounds and able to produce an intended result in four different qualities for other researchers. Intelligent model also able to overcome the traditional method in terms of time-consuming and consistency.

2. THEORETICAL WORK
2.1. Stepwise regression model

Stepwise regression analysis is a multiple regression analysis method. Multi-stepwise regression analysis is the most reliable mathematical statistical method in scientific research [19], [20] which can sort and analyze quantitative dependence between one dependent variable and multiple independent variables. Regression analysis is used to study the interdependence of multiple variables while stepwise regression analysis is frequently used to discover the ideal or most appropriate regression model to study the interdependence of variables in more depth. The stepwise regression potentially capable of adding or deleting one variable at a time have been the favourable methods [21].

The approach is to introduce the independent variables into the regression equation one at a time based on their influence on the dependent variables [22]. Stepwise regression minimizes the independent variables, X by using two processes which are forward and backward algorithm [23], [24]. The first approach is a forward selection method. Simultaneously, the significance test of each introduced independent variable is performed, and non-significant independent variables are eliminated. The independent variable starts to add into the regression equation only when they are statistically significant. By repeating this process, the most important variable from among numerous independent variables is finally chosen and a regression equation (mathematical model) which reasonably reflects the relationship between independent variables and dependent variables is established [25]. If the variables are not statistically significant, they will be eliminated one by one. For stepwise regression, the F-test and p-value are commonly used as tested values by referring to the p-value [26]. In statistics, the standard significance of p-value is less than 0.05 [27].

3. METHOD
3.1. Data preparation

The sample of dataset of agarwood oil used in this research was obtained from the previous researcher and consist of 660 samples between four different qualities (low, medium low, medium high and
high qualities). Out of that, 210 samples are from low quality, 90 samples from medium low qualities, 30 from medium high qualities while another 330 samples are from high quality classes [13]. There are eleven compounds which are dihydrocollumellarin, γ-eudesmol, α-guaiene, β-agarofuran, 10-epi-γ-eudesmol, ar-curcumene, Ycadine, valerianol, α-agarofuran, alloaromadendrene epoxide and β-dihydroagarofuran. The best four out of eleven agarwood oil chemical compounds were chosen based on the boxplot analysis. The selected chemical compounds employed in stepwise regression are γ-Eudesmol, 10-epi-γ-eudesmol, β-agarofuran and dihydrocollumellarin as highlighted in Table 1. All simulations were carried out using MATLAB software version R2015a.

Table 1. The compilation from boxplot result based on the median of abundances (%) for each chemical compounds

<table>
<thead>
<tr>
<th>Chemical Compounds</th>
<th>High Quality</th>
<th>Medium High Quality</th>
<th>Medium Low Quality</th>
<th>Low Quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>γ-eudesmol</td>
<td>0.64</td>
<td>0.55</td>
<td>0.63</td>
<td>0.41</td>
</tr>
<tr>
<td>10-epi-γ-eudesmol</td>
<td>0.76</td>
<td>0.5</td>
<td>0.65</td>
<td>0.41</td>
</tr>
<tr>
<td>β-agarofuran</td>
<td>0.53</td>
<td>0.99</td>
<td>0.32</td>
<td>0.01</td>
</tr>
<tr>
<td>dihydrocollumellarin</td>
<td>0.63</td>
<td>0.53</td>
<td>0.01</td>
<td>0.63</td>
</tr>
<tr>
<td>α-guaiene</td>
<td>0.26</td>
<td>0.01</td>
<td>0.17</td>
<td>0.01</td>
</tr>
<tr>
<td>ar-curcumene</td>
<td>0.5</td>
<td>0.01</td>
<td>0.01</td>
<td>0.76</td>
</tr>
<tr>
<td>β-dihydro agarofuran</td>
<td>0.33</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>Y-cadine</td>
<td>0.01</td>
<td>0.18</td>
<td>0.23</td>
<td>0.41</td>
</tr>
<tr>
<td>α-agarofuran</td>
<td>0.01</td>
<td>0.27</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>allo aromadendrene epoxide</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.36</td>
</tr>
<tr>
<td>Valerianol</td>
<td>0.35</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
</tr>
</tbody>
</table>

3.2. Flowchart of experimental set-up

Firstly, the experiment starts with empty variables as mentioned on flowchart in Figure 1. Then, the p-value will be computed if there are no variables in the stepwise regression model. There are two processes involved for stepwise regression which are forward selection and backward elimination. The observed p-value at each of the independent variables of agarwood oil were in detail. The value of p-value should be less than 0.05 for the X independent variable can be added into the model or otherwise the X variable is not added but the forward selection process will remain continuous until there is X variable in the regression model. Next, backward elimination will observe the p-value of X variables that have more than 0.05. An action of removing the variable X will be done or else, the selected variables will be maintained in the model. The selected variables will be the output feature of regression model and will be the input feature or marker to other intelligent models. (1)-(4) show the proposed calculation for the degree of freedom, coefficient of determination, adjusted R² and root mean square error (RMSE).

\[
\text{Degree of Freedom, } DF = P - Q
\]  

\[
\text{Coefficient of determination, } R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}
\]  

\[
\text{Adjusted } R^2 = 1 - \left( \frac{P - 1}{P - Q} \right) \times \frac{SSE}{SST}
\]  

\[
RMSE = \sqrt{\frac{SSE}{DF}}
\]

Where,
P = the number of rows from the data of agarwood oil samples
DF = error degrees of freedom
Q = number of coefficients
R² = coefficient of determination
SSR = sum of squared regression
SST = sum of squared total
SSE = sum of squared errors
RMSE = the estimation of the standard deviation of the error distribution
4. RESULTS AND DISCUSSION

In this section, the result obtained by training the regression model into four qualities which are low, medium low, medium high and high quality of agarwood oil samples. As specified in the methodology section, 660 samples of agarwood oil were employed in the analytical procedure. The data samples were then evaluated and developed into a stepwise regression model.

4.1. Generate stepwise regression

As can be seen in Figure 2, the independent variables are X3, X2, X1 and X4 which refer to the compound β-agarofuran, 10-epi-ɤ-eudesmol, ɤ-Eudesmol and dihydrocollumellarin, respectively. All the independent variables have been selected by observing the p-value of each variable where all of them have p-value less than 0.05 significance level. The results of estimated coefficients of predicted output for agarwood oil has been done and tabulated in Table 2. It is found that, the highest p-value found at intercept X2 which is 4.6661x10^{-51}, while the lowest value p-value is 2.0551x10^{-06} which is at intercept X4.

Based on the summary findings summarizes in Table 3 indicate that R^2 value is 0.756 or 75.6%. Therefore, according to the theory, the correlation coefficient, R^2 should have value below 80% for the best fit of stepwise regression. There were five numbers of predictors where belong to intercept and selected four X variables. The overall P-value for the F-test was 3.27x10^{-199} < 0.05. Those four chemical compounds of four different agarwood oil qualities which are ɤ-Eudesmol, 10-epi-ɤ-eudesmol, β-agarofuran and dihydrocollumellarin have proved that they all passed the performance for stepwise regression.
Stepwise regression of agarwood oil significant chemical compounds  

(Step 1. Adding x3, FStat = 1090.2342, pValue = 9.4581127e-142
2. Adding x2, FStat = 248.8138, pValue = 2.691568e-47
3. Adding x1, FStat = 56.7161, pValue = 1.67277e-13
4. Adding x4, FStat = 22.958, pValue = 2.0551e-06

Figure 2. Stepwise regression output generated from the MATLAB command window

Table 2. Estimate coefficients of stepwise regression

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate value</th>
<th>Standard Error, SE</th>
<th>t-statistics</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>-1.196</td>
<td>0.094016</td>
<td>-12.722</td>
<td>2.7332 × 10⁻³³</td>
</tr>
<tr>
<td>X1</td>
<td>0.49927</td>
<td>0.097009</td>
<td>5.1466</td>
<td>3.5111 × 10⁻⁰⁷</td>
</tr>
<tr>
<td>X2</td>
<td>2.5986</td>
<td>0.15815</td>
<td>16.431</td>
<td>4.6661 × 10⁻⁵¹</td>
</tr>
<tr>
<td>X3</td>
<td>2.3084</td>
<td>0.11077</td>
<td>20.84</td>
<td>2.2078 × 10⁻⁷⁴</td>
</tr>
<tr>
<td>X4</td>
<td>0.39217</td>
<td>0.081858</td>
<td>4.7909</td>
<td>2.0551 × 10⁻⁰⁶</td>
</tr>
</tbody>
</table>

Table 3. Summary output of stepwise regression

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of coefficients, Q</td>
<td>5</td>
</tr>
<tr>
<td>Degrees of freedom, DF</td>
<td>655</td>
</tr>
<tr>
<td>Root mean squared error, RMSE</td>
<td>0.671</td>
</tr>
<tr>
<td>R²</td>
<td>0.756</td>
</tr>
<tr>
<td>P-value</td>
<td>3.27 x 10⁻¹⁹⁹</td>
</tr>
<tr>
<td>SSR</td>
<td>915.9689</td>
</tr>
<tr>
<td>SST</td>
<td>1.2109 x 10⁸⁷</td>
</tr>
<tr>
<td>SSE</td>
<td>294.9402</td>
</tr>
</tbody>
</table>

Preliminary Linear regression equation model is as (5):

\[ Y \sim 1 + X1 + X2 + X3 + X4 \] (5)

hence, \( Y = -1.196 + 0.49927X1 + 2.5986X2 + 2.3084X3 + 0.39217X4 \)

5. CONCLUSION

Various research has attempted to demonstrate that the insufficiency of quality classification using conventional techniques might have an impact on the grading system. This paper effectively showed the development of the pre-analysis of agarwood oil classification using stepwise regression. This study has successfully applied a stepwise regression model to predict the agarwood oil quality classification. Four selected compounds are selected as input data which contain low, medium low, medium high and high quality of agarwood oil. With that, 0.756 of correlation coefficient, \( R² \) for the model result have been achieved. The findings strongly proved that agarwood oil obtained a best fit stepwise regression with the value of \( R² \) exactly below 0.8. Overall P-value also below 0.05 which 3.27 x 10⁻¹⁹⁹ in more specific. The findings can be evaluated and used as a reference in classifying the agarwood oil quality grading in the future.

ACKNOWLEDGEMENTS

The authors would like to express the gratitude to School of Electrical Engineering, Universiti Teknologi MARA (UiTM), Cawangan Johor, Kampus Pasir Gudang and 600-RMC 5/3/GPM (030/2022) for providing the financial support throughout this research. Also, to BioAromatic Research Centre of Excellence (BARCE), Universiti Malaysia Pahang (UMP), Forest Research Institute Malaysia (FRIM) and Advanced Signal Processing Research Group (ASPRG), Universiti Teknologi MARA for their support at various stages for this study.

REFERENCES

BIOGRAPHIES OF AUTHORS

Siti Mariatuil Hazwa Mohd Huzir was born in Malaysia who is currently pursuing her studies as a postgraduate student majoring in Electrical Engineering at UiTM Cavanggan Johor, Kampus Pasir Gudang. She received her B. Eng (Hons) of Electronics Engineering from Universiti Teknologi MARA (UiTM) Shal Alam Malaysia. She can be contacted at email: mariatuhazwah@gmail.com.
Stepwise regression of agarwood oil significant chemical compounds ... (Siti Mariatul Hazwa Mohd Huzir)

Aqib Fawwaz Mohd Amidon was born in Malaysia, on September 1996. He received his B. Eng. (Hons) of Electronic Engineering from Universiti Teknologi MARA (UiTM). He is currently a Software Engineer at Greatech Technology Berhad and at the same time as full time postgraduate students at School of Electrical Engineering, College of Engineering, Universiti Teknologi MARA, UiTM Shah Alam, Malaysia. He can be contacted at email: aqibfawwaz.080996@gmail.com.

Anis Hazirah ‘Izzati Hasnu Al-Hadi was born in Malaysia, on July 1998. She received her B. Eng (Hons) of Electronic Engineering from Universiti Teknologi MARA (UiTM). Currently, she is a postgraduate student at School of Electrical Engineering, College of Engineering, Universiti Teknologi MARA (UiTM) Shah Alam Malaysia. She can be contacted at email: anshzrh@gmail.com.

Ir. Ts. Dr. Nurlaila Ismail received her PhD in Electrical Engineering from Universiti Teknologi MARA, Malaysia. She is currently a senior lecturer at School of Electrical Engineering, College of Engineering, Universiti Teknologi MARA, Malaysia. Her research interests include advanced signal processing and artificial intelligence. She can be contacted at email: nurlaila0583@uitm.edu.my.

Ts. Dr. Zakiah Mohd Yusoff is a senior lecturer who is currently working at UiTM Pasir Gudang. She received the B. ENG in Electrical Engineering and PhD in Electrical Engineering from UiTM Shah Alam, in 2009 and 2014, respectively. In May 2014, she joined UiTM Pasir Gudang as a teaching staff. Her major interests include process control, system identification, and essential oil extraction system. She can be contacted at email: zakiah9018@uitm.edu.my.

Profesor Madya Dr. Saiful Nizam Tajuddin received his PhD from Universiti Malaysia Pahang (UMP). He is an Associate Professor and director of Bioaromatic Research Center of Excellence (BARCE) at Universiti Malaysia Pahang. He is a director and researcher at Synbion Sdn Bhd, Kuantan, Pahang, Malaysia. He has been a very active researcher and over the years had author and/or co-author many papers published in refereed journals and conferences. He can be contacted at email: saifulnizam@ump.edu.my.

Prof. Ir. Ts. Dr. Haji Mohd Nasir Taib received his PhD from UMIST, UK. He is a Senior Professor at Universiti Teknologi MARA (UiTM). He heads the Advanced Signal Processing Research Group at the School of Electrical Engineering, College of Engineering, UiTM. He has been a very active researcher and over the years had author and/or co-author many papers published in refereed journals and conferences. He can be contacted at email: dr.nasir@uitm.edu.my.