Journal of Fundamental and Applied Sciences

ISSN 1112-9867

Available online at http://www.jfas.info

Research Article

Special Issue

STATISTICAL ANALYSIS OF AGARWOOD OIL COMPOUNDS IN DISCRIMINATING THE QUALITY OF AGARWOOD OIL

N. S. A. Zubir[1,*], M. A. Abas[1], N. Ismail[1], M. H. F. Rahiman[1], S. N. Tajuddin[2] and M. N. Taib[1]

[1]Faculty of Electrical Engineering Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia

[2]Bio Aromatic Research Centre of Excellent, Universiti Malaysia Pahang, 26300 Gambang, Pahang, Malaysia

Published online: 05 October 2017

ABSTRACT

Enhancing and improving the discrimination technique is the main aim to determine or grade the good quality of agarwood oil. In this paper, all statistical works were performed via SPSS software. Two parameters involved are abundance of compound (%) and quality of the agarwood oil either low or high quality. The result showed that, there is clear significant normality test depends on the parameters through the boxplot and Quantile-Quantile (Q-Q) plot for data distribution. Then, Kolmogorov-Smirnov (K-S), non-parameter test and hypothesis were covered in normality test. The techniques proved their capabilities in statistical analysis for agarwood oil compounds and confirmed that the data is suitable for further work especially for classification.

Keywords: agarwood oil; SPSS; boxplot; Quantile-Quantile (Q-Q) plots; Kolmogorov-Smirnov (K-S); non-parameter test and hypothesis.

Author Correspondence, e-mail: nurulshakilaaz@gmail.com

doi: http://dx.doi.org/10.4314/jfas.v9i4s.3

1. INTRODUCTION

Agarwood oil is one of the impregnated heartwood of the Aquilaria species which part of the Thmelaeaceae family. Agarwood oil is extracted from the agarwood oil tree as known as gaharu tree in Malaysia. Agarwood oil is one of the business sectors that increase the economic. This is because the demand for agarwood oil is very high, especially in the Middle East. Normally, in the Middle East agarwood oil is main choices for wedding ceremony. People in Middle East burn down agarwood oil during wedding ceremony and believe that the custom as a symbol of wealth. Other than that, agarwood oil is used as traditional medical preparation and perfume [1-6].

System classification in system identification is one of the techniques that can be used to classify data scientifically to different classes according to certain constraints. Besides that, statistical knowledge is rooted as an origin in system identification.

Statistical analysis is one of the crucial parts of the research process that become suggested method to identify the type of the data and analyze the data especially for developing the classification models. In this part, the analysis involved of preliminary analysis and statistical technique to explore relationships among variables and comparison of the group. In preliminary analysis, data have to be prepared for error performance checked. The analysis consists of descriptive statistics and graphs, the manipulation of data and checking the reliability of the scales. Besides that, the major statistical technique that commonly used to discovered the relationships are correlation, partial correlation, multiple regression, logistic regression and factor analysis.

2. THEORETICAL BACKGROUND

2.1. Data Distribution

2.1.1. Boxplot

Boxplot is known as the effective and efficient visualization alternative. Boxplot are begin with sorted the scores. In Fig. 1, there are four sized of group made from ordered scores consist of median, upper quartile and lower quartile and whisker. Each score have 25% place in each group. The line in the boxplot was dividing in the groups called as quartiles and the

groups are starting from the bottom. The function of median (middle quartile) is to marks the mid-point of the data and median normally has two conditions, which are half the scores are greater or equal to that value or half are less.

The interquartile range at the middle box represents the 50% scores for the group. Interquartile be as a reference for ranges of lower to upper quartile. The third part is lower and upper quartile which is 25%, and the last is whisker which shows the scores outside the middle 50% [7-8].

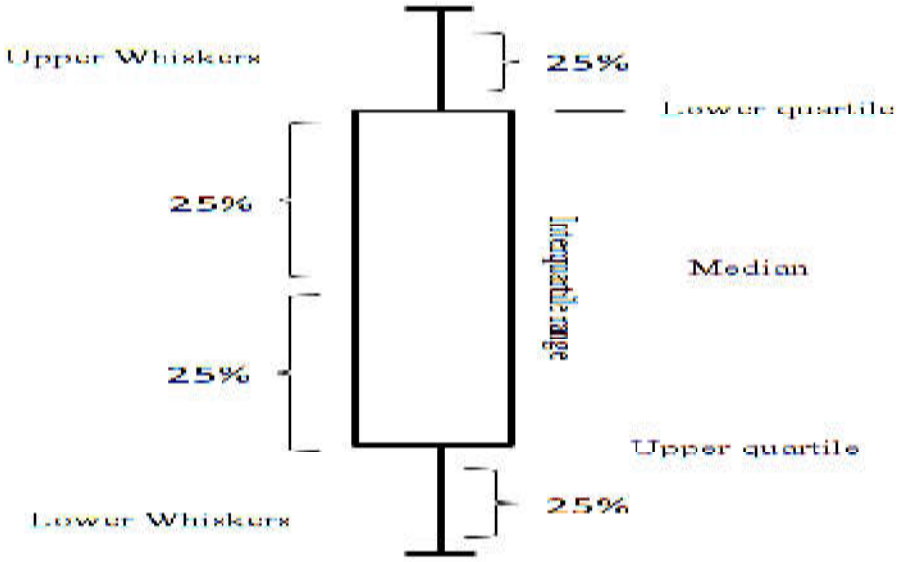


Fig.1. Boxplot structure

2.1.2. Quantile-Quantile (Q-Q) Plots

The Q-Q plot is used to evaluate the normality of data in terms of graphical assessment. The straight line function as a reference of data points either closer or near that identified as a normal data. This method consists of two vertical which is vertical axis and horizontal axis. The vertical axis is for ordered response value and the horizontal axis represent as median normal order statistics. The conclusion of Q-Q plot will be observing trough the points that assembling near the straight line. Otherwise, the not normal condition is when the scattered

points lying further from the straight line and it is called as outliers [9-10].

2.2. Normality Test

2.2.1. Kolmogorov-Smirnov (K-S)

Normality test is the important part required to observing the distribution of the data. Normality test have two condition which are depends on the data distribution either normal or not normal. If the data distribution was normal, the parametric method can be applied in analyzing the data. Otherwise, if the data is not normal the non-parametric metric should be applied for analyzing the data. Besides that, Kolmogorov-Smirnov (K-S) was used to indicate the category of the data distribution depends on the significance p-value. The p-value function is to compare between cumulative distributions with expected cumulative normal distribution of the data. If the value of p-value larger than 0.05, the data indicates as a normal distribution other than that it categorized as not normal [10].

2.2.2. Non-Parametric Test

The non-parametric is selected depends on the null hypothesis. It is used to analyze data after accessing the distribution of data. The non-parametric statistic is applied when the data is not normal distributed [11]. Besides that, non-parametric is known as distribution free test because the distribution data do not able to assume.

2.2.3. Hypothesis Test

Hypothesis testing is defined as a summary to determine the suitable type of statically method. The hypothesis testing can denied the testing claim, analyze the intensity of the data either support or reject the conclusion or in terms of population parameter. Other than that, hypothesis testing has two types which are null hypothesis or alternative hypothesis. Both types are refers to the significant value of data and in this study, alternative hypothesis based on data statistical.

3. METHODOLOGY

Fig. 2 illustrates the hierarchy structure that used to implement the statistical of agarwood oil in this study. The analysis works starts with agarwood oil data collection consist of two parameters, which are abundance of compound (%) and quality of the agarwood oil either low

or high quality. After that, this works followed by data distribution of parameters which are boxplot and Q-Q plot. Both methods were done to have general impression on the distribution of agarwood oil data. Then, normality test was performed to test data by using 3 methods, Shapiro-Wilk test, non-parameter test and hypothesis. Lastly, summary of statistical analysis was determined.

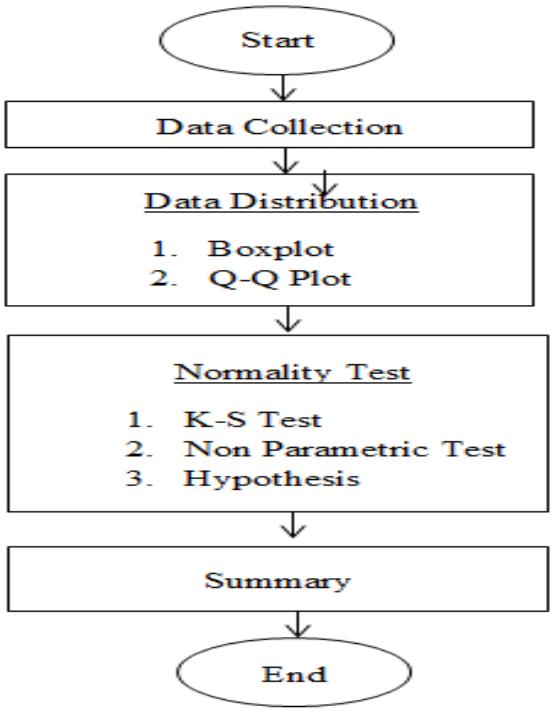


Fig.2. Hierarchy structure of statistical approach

3.1. Data Collection

The agarwood oil data collection in this study was obtained at Forest Research Institute Malaysia (FRIM) and University Malaysia Pahang (UMP) [12], which is consist of gas chromatography-mass spectrometry (GC-MS) data. The chemical compound of agarwood oil are either high or low quality. There are seven significant chemical compounds of sixteen samples of agarwood oils. The significant compounds has been used as input and quality of the agarwood oil either high or low has been used as an output to the classification system.

B. Data Distribution

3.1.1. Boxplot

To start creating the boxplot it is by entering the agarwood oil data into two columns which are 1) types of compound that consist of 7 compounds as independent variables (x-axis) and 2) abundance of compound (%) as a dependent variable (y-axis). The appropriate range and

nominal variable was set. In SPSS version 12, at the graphs menu, the 'boxplot' was selected and the user just rename of the x-axis and y-axis. Then, the performance of boxplot was observed.

3.1.2. Quantile-Quantile (Q-Q) Plots

To plot Quantile-Quantile (Q-Q), it is started with entering the data file. In the graph menu, Q-Q plot was selected and the 'analyze description' was chosen. When the little dots were at the line pretty well or near the straight line with some random scatter, the data classified as normal distribution. The variables and title was renamed in the box given.

3.2. Normality Test

In normality test, the steps are same with the data distribution which is load the data and place in the 2 column. The data was analyzing by select the descriptive analysis and explore. The new window will be pop up. The data will be transfer to the dependent and independent list. The new window will be pop up again to double check the descriptive and click plot. Kolmogorov-Smirnov (K-S), non-parameter test and hypothesis will be display. The test statistic was shown in the Table 1. Kolmogorov-Smirnov (K-S) was been chosen because the data more than 50 dataset. The P-value is less than 0.05. That is mean reject the null hypothesis and conclude that the data from a not normal data distribution.

4. RESULTS AND DISCUSSION

4.1. Data Collection

Table 1, 2 and 3 show the selection of the optimal number of thresholds to evaluate the performance based on the value accuracy of training, validation and testing data set of agarwood oil. The optimal number of threshold would be identified the group qualities of agarwood oil either high (1) or low (0). The numbers of thresholds was adjusted in the ranged of 0.1 to 1. An optimal threshold (or set of 0.5 thresholds) was selected based on the value of 100 % accuracy on the training, validation and testing data set prove in Fig. 3. The low value of agarwood oil from the optimal number of threshold was discriminate as a low quality (0). Otherwise, for the agarwood oil that has higher value than optimal number of threshold was discriminate as high quality (1).

Table 1. Confusion matrix for training data set

Number of Thresholds	Sensitivity (%)	Specificity (%)	Predictive (%)		Accuracy (%)
			Positive	Negative	
0.1	78.6	100	100	94.7	95.6
0.2	78.6	100	100	94.7	95.6
0.3	78.6	100	100	94.7	95.6
0.4	78.6	100	100	94.7	95.6
0.5*	100	100	100	100	100
0.55*	100	100	100	100	100
0.6*	100	100	100	100	100
0.7	100	98.1	93.3	100	98.5
0.8	100	81.5	58.3	100	85.3
0.9	100	81.5	58.3	100	85.3
1.0	100	0.0	20.6	NaN	20.6

*Highest value of accuracy

Table 2. Confusion matrix for validation data set

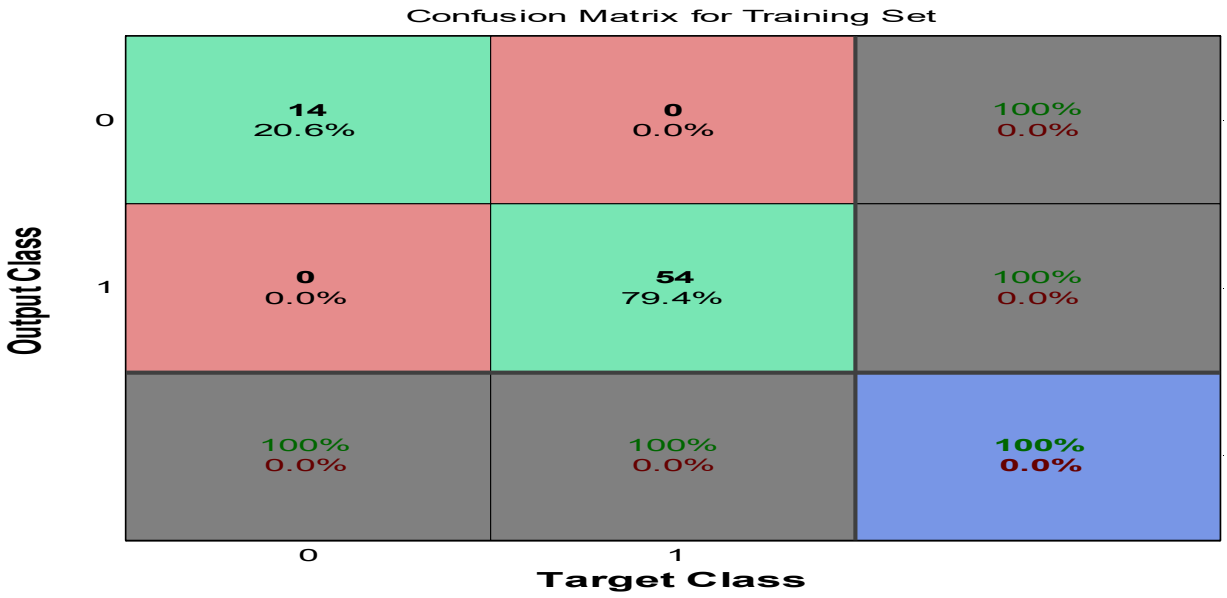
Number of Thresholds	Sensitivity (%)	Specificity (%)	Predictive (%)		Accuracy (%)
			Positive	Negative	
0.1	50	100	100	92.3	92.9
0.2	50	100	100	92.3	92.9
0.3	50	100	100	92.3	92.9
0.4	50	100	100	92.3	92.9
0.5*	100	100	100	100	100
0.55*	100	100	100	100	100
0.6*	100	100	100	100	100
0.7	100	100	100	100	100
0.8	100	100	100	100	100
0.9	100	100	100	100	100
1.0	100	0.0	14.3	NaN	14.3

*Highest value of accuracy

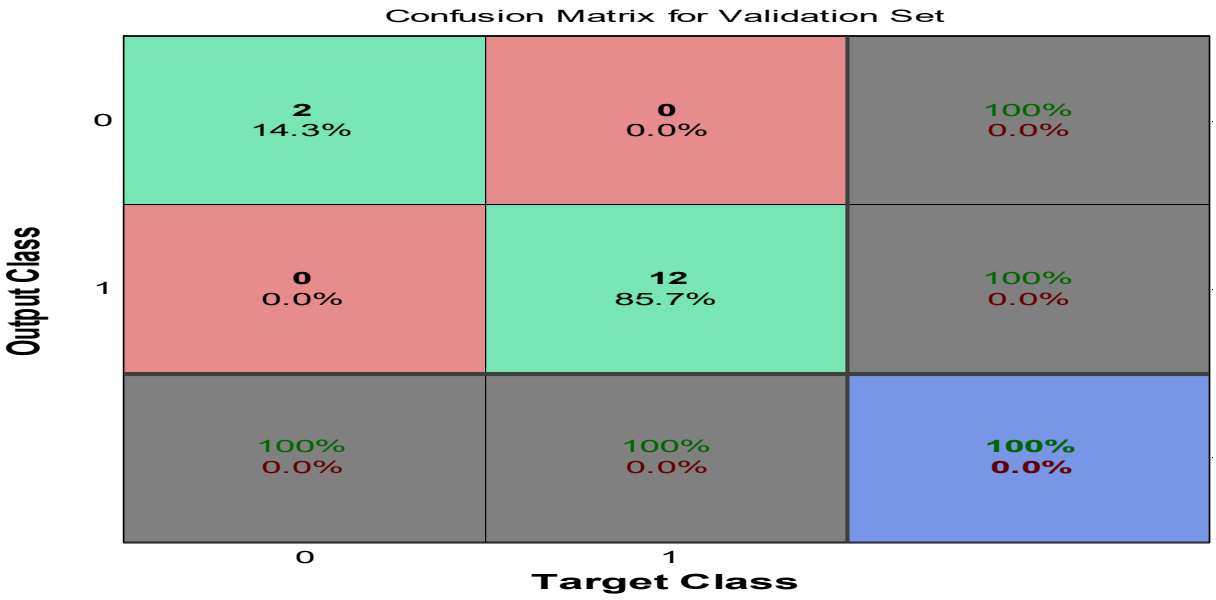
Table 3. Confusion matrix for testing data set

Number of Thresholds	Sensitivity (%)	Specificity (%)	Predictive (%)		Accuracy (%)
			Positive	Negative	
0.1	0.0	100	NaN	85.7	85.7
0.2	0.0	100	NaN	85.7	85.7
0.3	0.0	100	NaN	85.7	85.7
0.4	0.0	100	NaN	85.7	85.7
0.5*	100	100	100	100	100
0.55*	100	100	100	100	100
0.6*	100	100	100	100	100
0.7	100	100	100	100	100
0.8	100	83.3	50	100	85.7
0.9	100	83.3	50	100	85.7
1.0	100	0.0	14.3	NaN	14.3

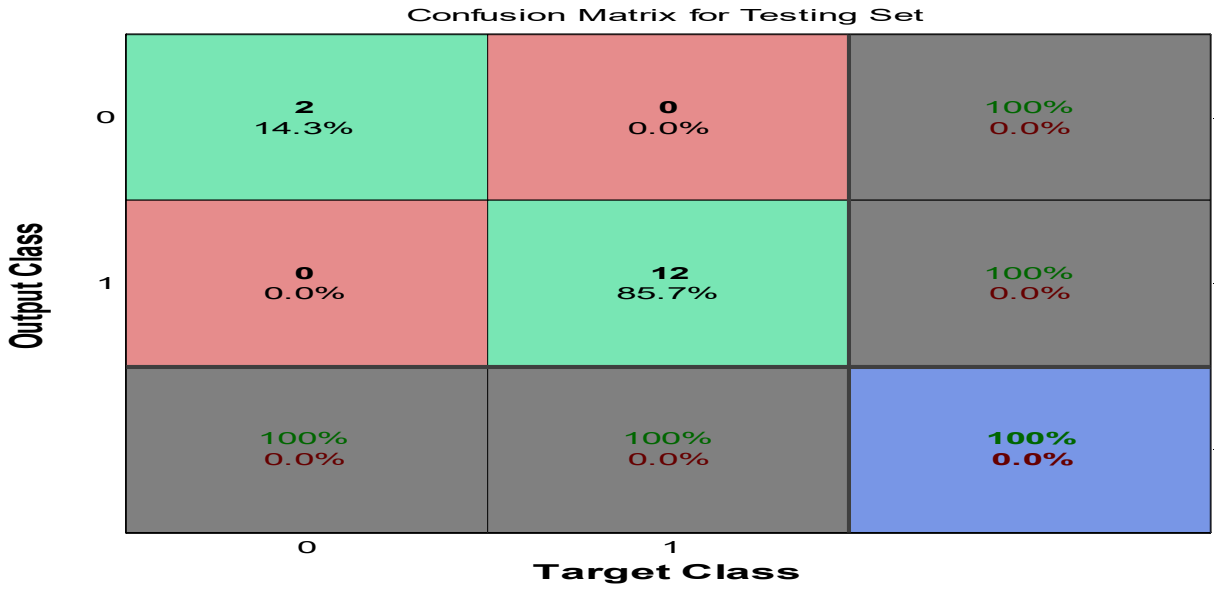
*Highest value of accuracy



(a)



(b)



(c)

Fig.3. Confusion matrix for (a) training (b) validation (c) testing data set using 0.5 thresholds

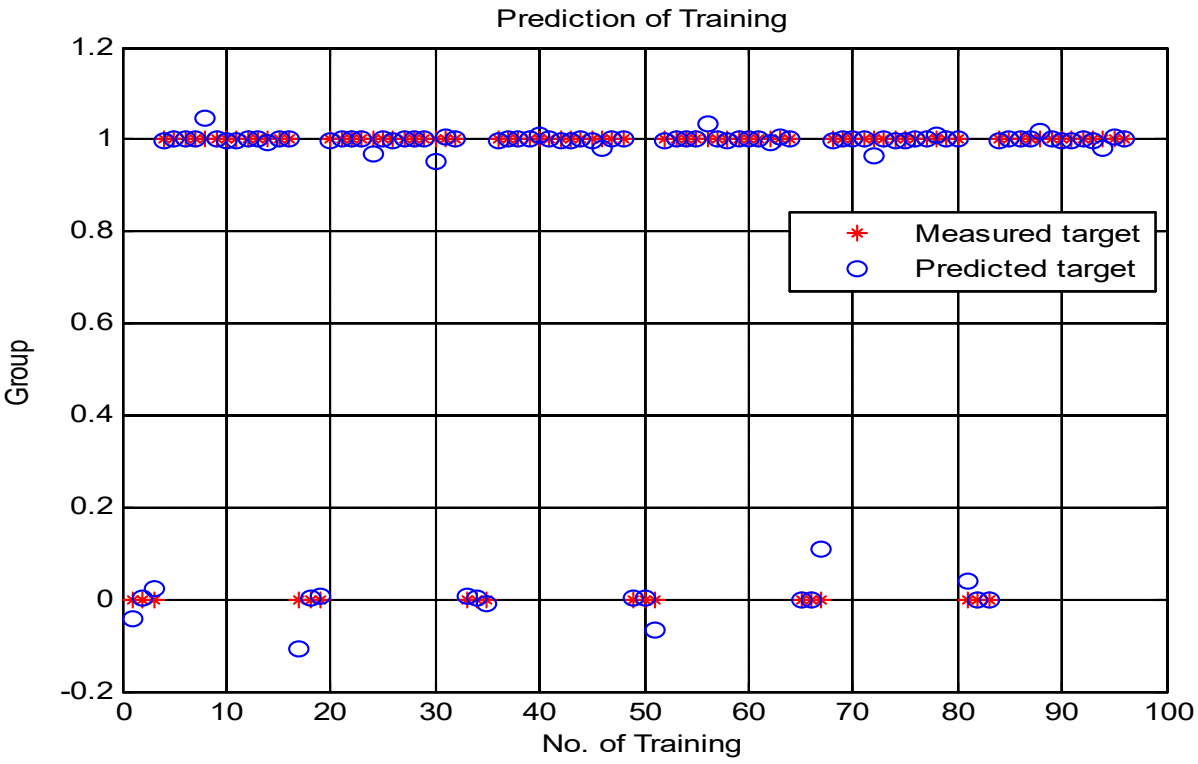


Fig.4. Threshold of training network

Fig. 4 shows the measured and predicted target for threshold of training network. The low and high quality represented by 0 and 1 respectively. 0.5 thresholds indicated high quality agarwood oil more than low quality. The predicted target presented high accuracy result as compared to measured target.

4.2. Data Distribution

4.2.1. Boxplot

Fig. 5 shows that the boxplot with using data consist of 7 compounds and the abundances (%) which are C1 = β-agarofuran, C2 = α-agarofuran, C3 = 10, -epi-ϒ-eudesmol, C4 = ϒ-eudesmol, C5 = longifolol, C6 = hexadecanol and C7 = eudesmol. Compound 1, 2 and 3 are significantly present in high quality and compound 5, 6 and 7 are significantly present in low quality. Compound 4 is neutral. It means this compound was found significantly in high and low qualities [13]. Compound 4 shows that their 50% quartile is located in the middle and their whiskers have similar distance and obviously there are no outliers. However, the compound 1, 2 and 3 show that the whiskers is longer than the other side and the quartile for upper and lower not achieve 25% each quartile and have outliers. In this case, the distribution is not normal. Same goes to compound 5, 6 and 7 the interquartile ranged is so small and there

are many outliers. Those compounds are also under not normal distribution. These observation proceed with another test which is Q-Q plots to make it clearly the conclusion.

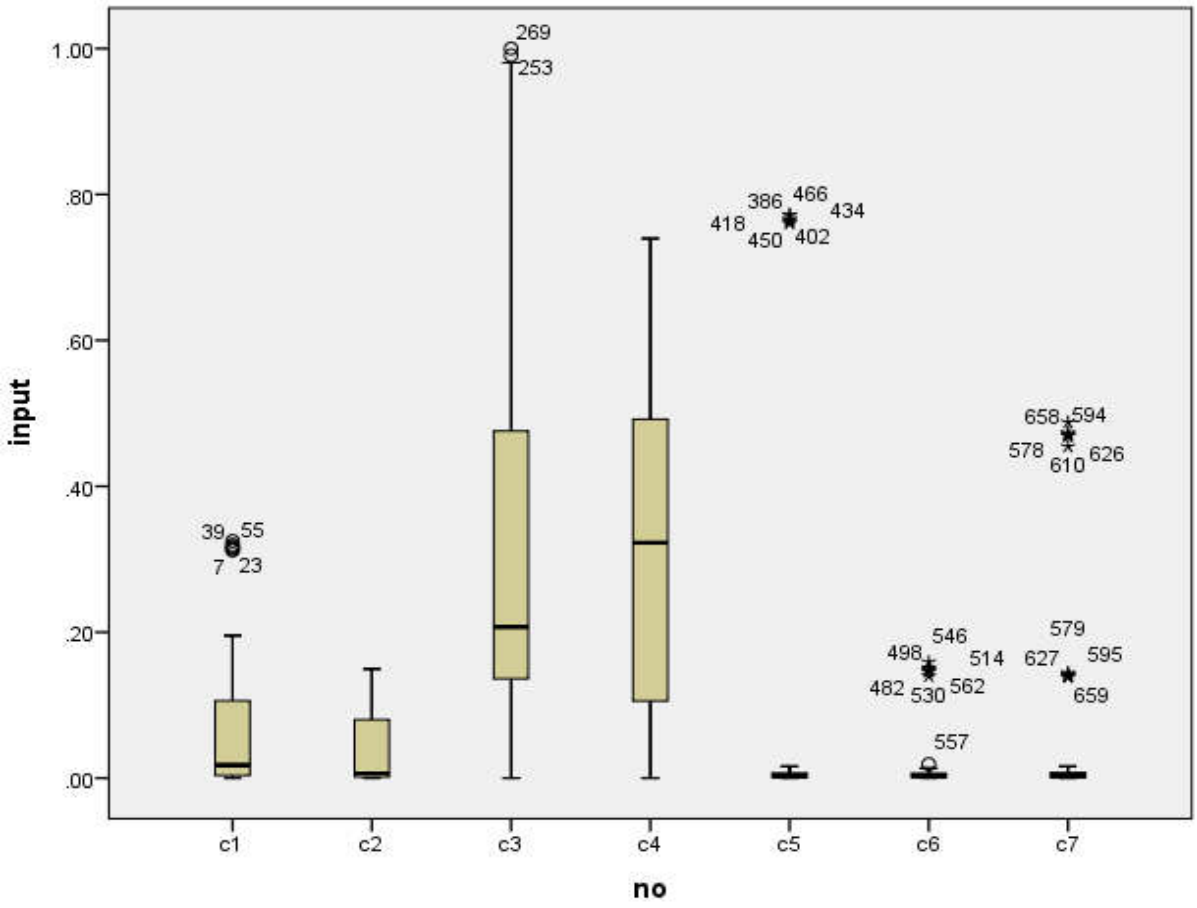


Fig.5. Boxplot

4.2.2. Quantile-Quantile (Q-Q) Plots

Fig. 6 to 12 show the Q-Q plot of 7 compounds used in this study which are C1 = β-agarofuran, C2 = α-agarofuran, C3 = 10, -epi-ϒ-eudesmol, C4 = ϒ-eudesmol, C5 = longifolol, C6 = hexadecanol and C7 = eudesmol. These data were represented as the non-normal data distribution. It is proved by the trends of agarwood oil data points diverged from the straight line for all figures. Specifically, it is observed that in Fig. 6 to Fig. 8, compound 1 to compound 3 has similar trend in Q-Q plots. It is notified that, the compounds 1 to 3 were belonging to high quality of agarwood oil. Next, compound 4 in Fig. 9 has scattered plot lying near the straight line which is close to the 45-degree reference line. It found that this compound is neutral which is this compound exist in high and low quality of agarwood oil. Then, for Fig. 10 to Fig. 12, the similar trend in Q-Q plots present and confirm

that these compound exist significantly in low quality of agarwood oil. These observations confirmed previous finding from other researcher [13].

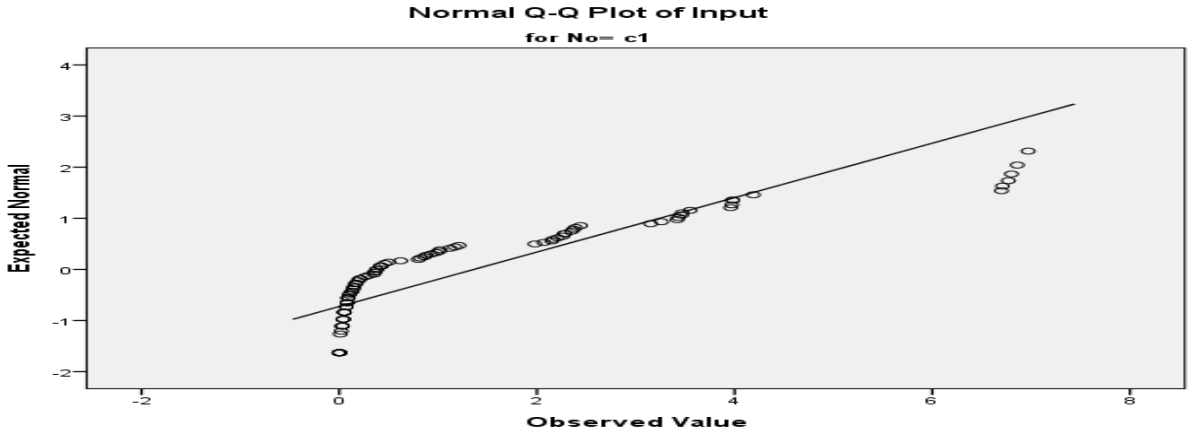


Fig.6. Q-Q plot compound 1

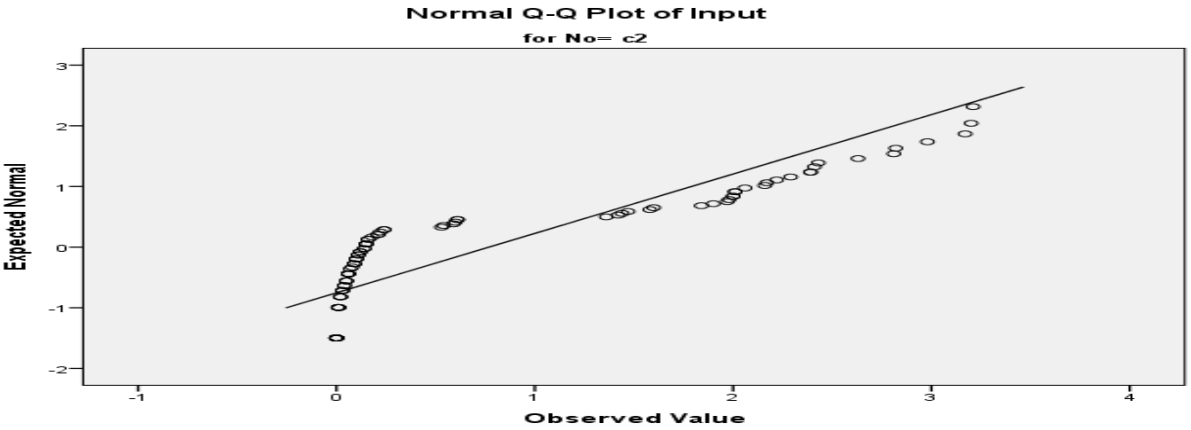


Fig.7. Q-Q plot compound 2

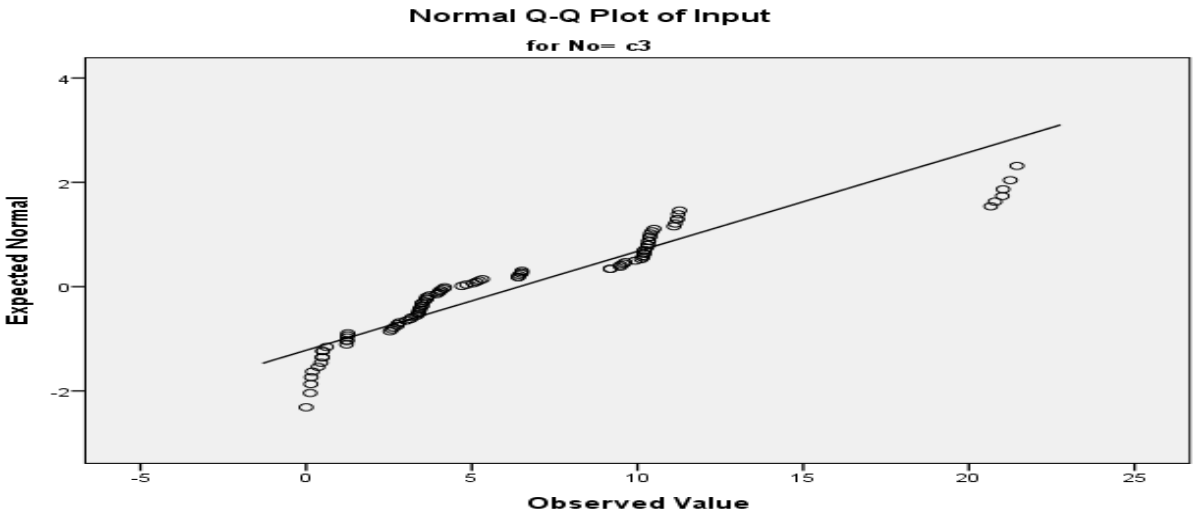


Fig.8. Q-Q plot compound 3

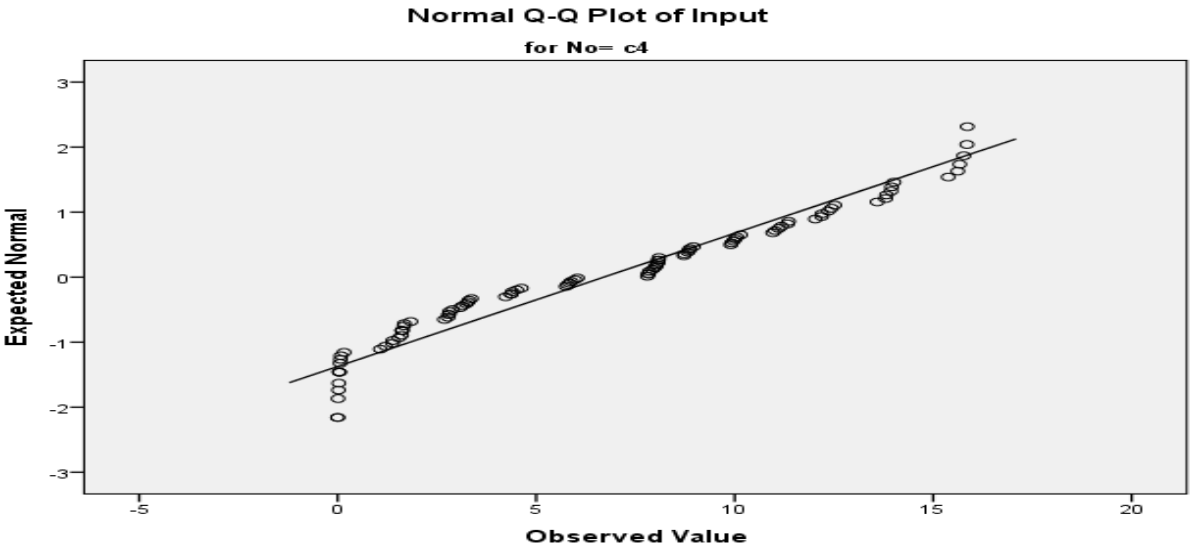


Fig.9. Q-Q plot compound 4

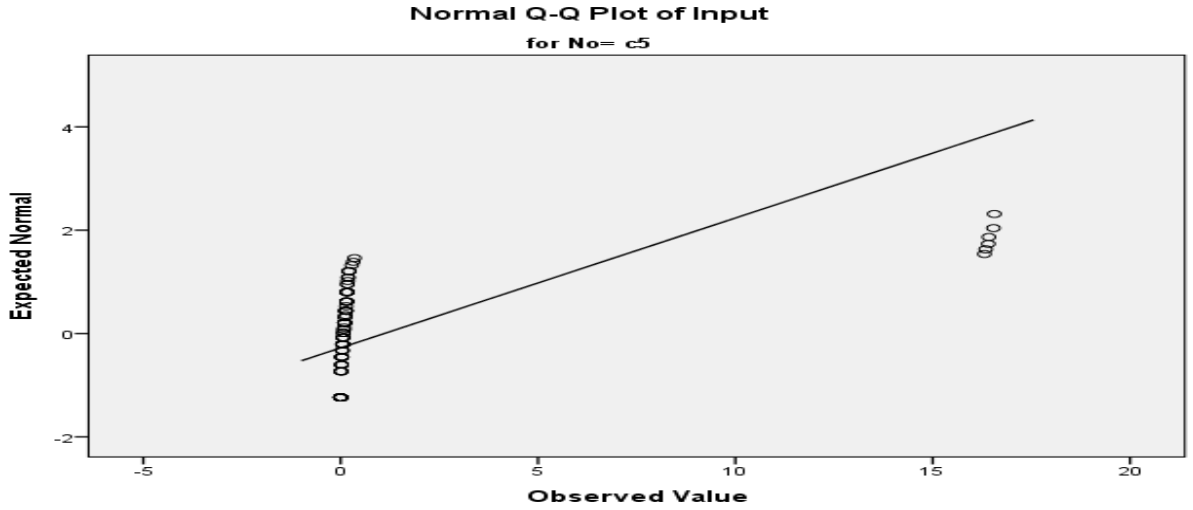


Fig.10. Q-Q plot compound 5

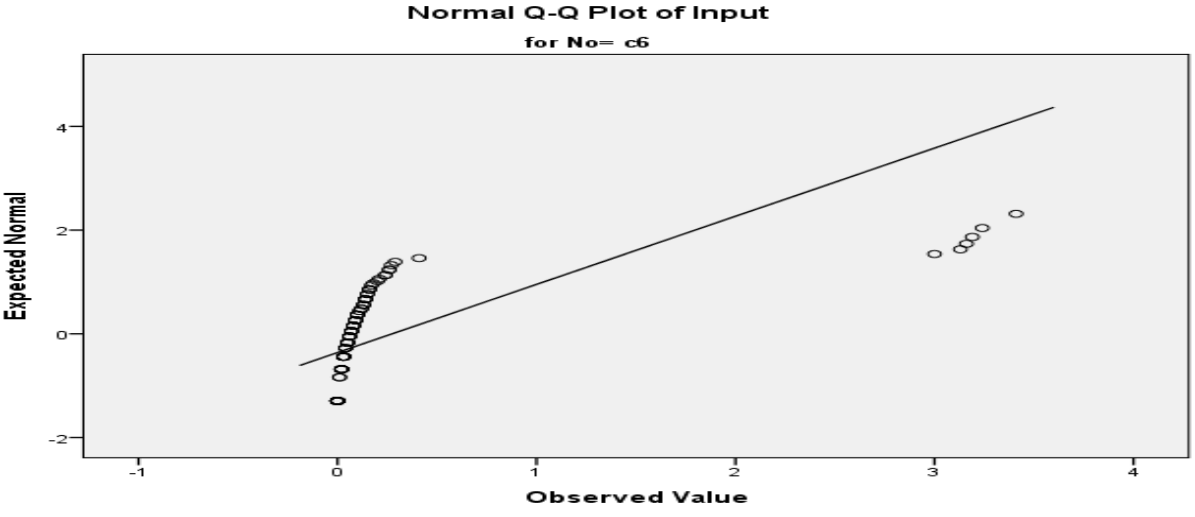


Fig.11. Q-Q plot compound 6

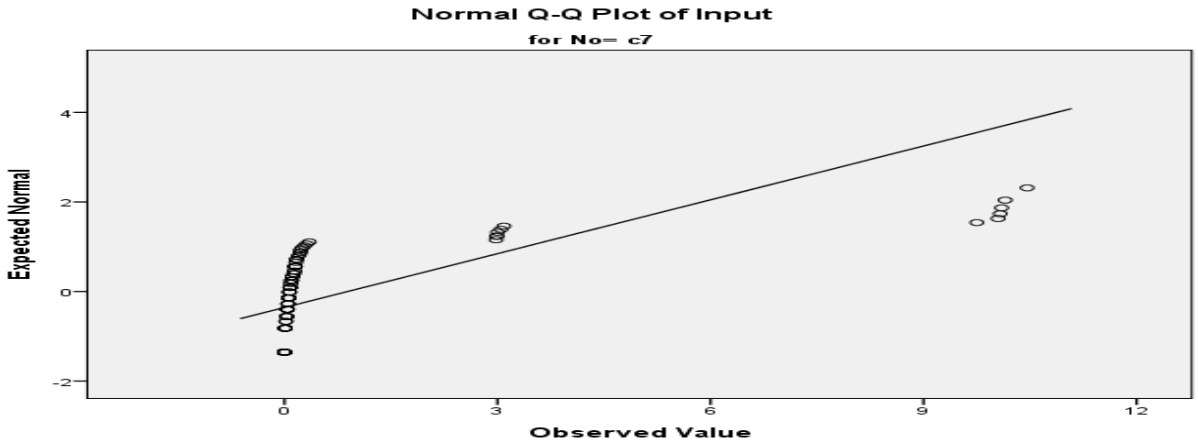


Fig.12. Q-Q plot compound 7

4.3. Normality Test

4.3.1. Kolmogorov-Smirnov (K-S) Test

Table 4 shows that the significant value as known as p-value is less than 0.05 and there is enough evidence to reject the null hypothesis. Regarding on t-value, df and significance value provided by SPSS, the result can show the population group either in the same or different population. P-value is select 0.05 depends on the standard alpha level [10]. T-value shows the degree of freedom of this agarwood oil data. Among 7 compounds, compound 4 have higher significant value which is 0.0001 other than another compound. Besides it also has been proven at box plot, compound 4 was found in high and low quality of agarwood oil.

Table 4. Kolmogorov-Smirnov test

Compound	Statistic	df	Sig.
1	0.239	96	0
2	0.323	96	0
3	0.266	96	0
4	0.127	96	0.001
5	0.513	96	0
6	0.419	96	0
7	0.461	96	0

4.3.2. Hypothesis Test

In Fig. 13 shows the decision at hypothesis test summary that the null hypothesis was rejected.

So that, it was classified as alternative hypothesis which is the hypothesis tends to population parameter assumption of null hypothesis is false. This method was support by the P-value as known as probability value for statistical evidence. The p-value of 0.05 was used as the significance probability (P-value) [13]. The value of significance level is below than 0.05 which us 0.00. This statement proves the data is non-parameter based on the hypothesis.

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of input is the same across categories of no.	Independent-Samples Kruskal-Wallis Test	.000	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

Fig.13. The hypothesis test summary

5. CONCLUSION

The preliminary analysis by using statistical technique, which is to explore the relationships among variables and comparison of the group was presented in this study. The works consist of boxplot, Q-Q plot and normality test (Kolmogorov-Smirnov (K-S), non-parametric and hypothesis. It showed that the data distribution was not normal and acceptable to be used for future work due to the result from boxplot and Q-Q plots distribution. Kolmogorov-Smirnov (K-S) Test proved that the data distribution is not normal and it is belongs to non-parametric. The hypothesis decision in this research was reject the null hypothesis.

6. ACKNOWLEDGEMENTS

The author would like to acknowledge PICO and Advanced Signal Processing (ASP) groups from Faculty of Electrical Engineering and staff from Faculty of Industrial Science and Technology (FIST) UMP for guidance as well as data collection. Also, to E-Science Grant (MOSTI) 02-01-16-SF0092 awarded to Associate Professor Dr. Saiful Nizam Tajuddin for financial support.

7. REFERENCES

[1] Ismail N, Rahiman M H, Taib M N, Ibrahim M, Zareen S, Tajuddin S N. A review on agarwood and its quality determination. In IEEE 6th Control and System Graduate Research Colloquium, 2015, pp. 103-108

[2] Hashim Y Z, Kerr P G, Abbas P, Salleh H M. Aquilaria spp.(agarwood) as source of health beneficial compounds: A review of traditional use, phytochemistry and pharmacology. Journal of Ethnopharmacology, 2016, 189:331-360

[3] Yang D, Wang J, Li W, Dong W, Mei W, Dai H. New guaiane and acorane sesquiterpenes in high quality agarwood "Qi-Nan" from Aquilaria sinensis. Phytochemistry Letters, 2016, 17:94-99

[4] Subasinghe S M, Hettiarachchi D S. Characterisation of agarwood type resin of Gyrinops walla Gaertn growing in selected populations in Sri Lanka. Industrial Crops and Products, 2015, 69:76-79

[5] Dahham S S, Tabana Y M, Hassan L E, Ahamed M B, Majid A S, Majid A M. In vitro antimetastatic activity of Agarwood (Aquilaria crassna) essential oils against pancreatic cancer cells. Alexandria Journal of Medicine, 2016, 52(2):141-150

[6] Ismail N, Rahiman M H, Jailani R, Taib M N, Ali N A, Tajuddin S N. Investigation of common compounds in high grade and low grade Aquilaria Malaccensis using correlation analysis. In IEEE Control and System Graduate Research Colloquium, 2012, pp. 277-281

[7] Mirzargar M, Whitaker R T, Kirby R M. Curve boxplot: Generalization of boxplot for ensembles of curves. IEEE Transactions on Visualization and Computer Graphics, 2014, 20(12):2654-2663

[8] Yusoff S, Wah Y B. Comparison of conventional measures of skewness and kurtosis for small sample size. In IEEE International Conference on Statistics in Science, Business, and Engineering, 2012, pp. 1-6

[9] Yang D, Qi H. An effective nonparametric quickest detection procedure based on QQ distance. In IEEE International Conference on Acoustics Speech and Signal Processing, 2010, pp. 3786-3789

[10] Ito M, Yoshida K, Hachiya H, Mamou J, Yamaguchi T. Quantification of the scatterer distributions for liver fibrosis using modified QQ probability plot. In IEEE International Ultrasonics Symposium, 2014, pp. 2394-2397

[11] Razali N M, Shamsudin N R, Maarof N N, Ismail A. A comparison of normality tests using SPSS, SAS and MINITAB: An application to health related quality of life data. In IEEE

International Conference on Statistics in Science, Business, and Engineering, 2012, pp. 1-6

[12] Najib M S, Ahmad M U, Funk P, Taib M N, Ali N A. Agarwood classification: A case-based reasoning approach based on E-nose. In IEEE 8th International Colloquium on Signal Processing and its Applications, 2012, pp. 120-126

[13] Ismail N, Rahiman M H, Taib M N, Ali N A, Jamil M, Tajuddin S N. Application of ANN in agarwood oil grade classification. In IEEE 10th International Colloquium on Signal Processing and its Applications, 2014, pp. 216-220

How to cite this article:
Zubir NSA, Abas MA, Ismail N, Rahiman MHF, Tajuddin SN, Taib M N. Statistical analysis of agarwood oil compounds in discriminating the quality of agarwood oil. J. Fundam. Appl. Sci., 2017, *9(4S), 45-61.*