eISSN 2550-1453 http://eproceedings.chemistry.utm.my/

Partial east squares correlation analysis of ethanol and ethanoic acid in vinegar using attenuated total reflectance-fourier transform infrared spectroscopy

‘Aqilah Fadhil and Hasmerya Maarof*
Department of Chemistry, Faculty of Science, Universiti Teknologi Malaysia, 81310 Johor Bahru, Malaysia
Corresponding Author: hasmerya@kimia.fs.utm.my

Article history :
Received 20 June 2017
Accepted 21 July 2017

ABSTRACT

Vinegar is a common material that is use in our daily life as it has vast usage, from being used in cooking and salad dressing, up to as cleaning substance due to its acidity behavior. Vinegar is made from different types of sources such as wine, apple, rice and others by fermentation of ethanol into ethanoic acid. Wine being the source of vinegar cause doubts in consumer when buying it in the market especially Muslims. According to Malaysian Food Act 1983, it is said that the content of acetic acid in any vinegar must not be more than 12.5 % w/v. In this study, Attenuated Total Reflectance-Fourier Transform Infrared Reflectance (ATR-FTIR) technique and Partial Least Squares Regression (PLS) method was used to build a model to quantify ethanol and ethanoic acid in standard samples and apply it to four commercialized vinegar. The standard samples are from 10 % v/v to 90 % v/v with 20 % v/v interval. Another four commercialized vinegar was used as real samples. Therefore, there are a total of 22 samples were scanned using ATR-FTIR and the absorbance values were recorded from 650 to 4000 cm^{-1}. The digitized spectra were saved in Microsoft Excel and later imported to MATLAB for further analysis. Then the spectrophotometric data were analyzed for PLS Regression. Each spectrum data was selected every 1cm^{-1} interval and data at Y-column 1592 which is 818 cm^{-1} was chosen as it gives the best result when building model. The model obtained shows the model can be accepted as the value of R^2 is 0.996, Root Mean Square Error of Calibration (RMSEC) of 0.0005535, Root Mean Square Error of Cross-Validation (RMSECV) of 0.0011474 before data of real samples were applied. After data of real samples were applied into this model, the R^2 obtained maintained high which is 0.999, while the values of Root Mean Square Error of Prediction (RMSEP) is 0.00040802 which is very small.

Keywords: Partial Least Squares, Correlation Analysis, Attenuated Total Reflectance, Principal Component Analysis, vinegar, ethanol, ethanoic acid

1. INTRODUCTION

Vinegar is one of the most common ingredients used in any food industry especially in fermentation and salad dressing. Some used it as preservation or flavoring in food and some even used it as cleaning agent in some cases due to its acidic behavior. What is vinegar? Vinegar is a solution that contains water, different trace materials which give different taste to the flavor and acetic acid which is from the fermentation of alcohol or more specifically ethanol (C_2H_5OH) into acetic acid or also known as ethanoic acid (CH_3COOH). The activity of vinegar bacteria or also known as acetic acid bacteria, convert ethanol into ethanoic acid by oxidation process as shown in the equation below [1]. The formation of ethanoic acid is crucial because ethanoic acid is the main ingredient of vinegar.

$$C_2H_5OH + O_2 \longrightarrow CH_3COOH + H_2O$$

Vinegar can be from different types of sources such as wine, apple, distilled alcohol, and many others. Different sources give different benefits and flavor. However, the core ingredient which is acetic acid what makes it a vinegar. The conversion of wine or any fruit juices into vinegar is a chemical process that occurs naturally by time. In this process, the ethanol undergoes partial oxidation which forms of acetaldehyde. There are three stages involved. During the first stage, the sugars from wine or fruit juices are broken down in the absence of oxygen (O_2) by yeast which gives results to the formation of ethanol and carbon dioxide (CO_2). In the second stage, the addition of oxygen enables the acetic acid bacteria or vinegar bacteria to produce amino acid, water and other compounds. Lastly, in the third stage, the acetaldehyde which is from the partial oxidation of ethanol is transformed into acetic acid that is the key ingredient in vinegar [1].

In today’s world, every tiny information will spreads fast. So fast that it is nearly impossible to get rid of the rumors or issues even if it is not true. The issues of *Halal* food has been growing bigger day by day. This shows how much concern the community have regarding this matter. This is because a lot of companies in vinegar market has taken an advantage of the naivety of society in Malaysia where many thought as long as there is *Halal* logo on the wrapping, then it is safe. The concern about *Halal* status of vinegar is due to the nature of how it is made. The source of vinegar which is wine is worrisome to those who does not know how chemical changes work. Yes, there are indeed some of the vinegar that put wine as one of the ingredients on the list like Balsamic vinegar[1]. But, the reason why the company listed wine as one of the ingredients is either because of its originality which is from wine or because wine was added after the final production of the vinegar. In other words, after vinegar is produced, the company added wine to add more flavor into it [2]. The latter case may not be always true but it is more often found in high end gourmet. However, vinegar should contain at least 4 % w/v acetic acid as required for preserving and prickling as stated in Food Regulation 1985 [3].

The purpose of this study is to build a model to quantify ethanol in synthetic sample using attenuated total reference (ATR-FTIR) and to apply model to four different types of vinegar. However, this study does not include homemade vinegar and it is limited to analysis using ATR-FTIR technique only. The nature of how the vinegar is made should not be a cover for all the vinegar-making company to use 'wine' as ingredient carelessly. This wrong business etiquette affects the whole vinegar industry. Therefore, the benefit of building this model is it can help the vinegar industry in the near future by verifying the alcohol content in a vinegar before sending it to the market.

2. EXPERIMENTAL

The experiment was divided into four main stages. The first stage is the sample preparation of different concentrations of standard samples of ethanol and ethanoic acid also sample preparation for real samples. 95 % v/v of ethanol and glacial acetic acid were used as stock solution. The second stage is acid-base titration for standard samples of ethanoic acid at different concentration of 10 % v/v, 30 % v/v, 50 % v/v, 70 % v/v and 90 % v/v with sodium hydroxide together with phenolphthalein as indicator [4]. The third stage is analysis of all samples with ATR-FTIR. A total of 22 samples were analysed together with their replicates. All spectra data was obtained in the form of transmittance which will be converted into absorbance later on. The last stage is PLS regression model built up. MATLAB software was used to build up this model using all of the data that have been obtained.

3. RESULTS AND DISCUSSION

3.1. ATR-FTIR Analysis

There were a total of 22 samples were analyzed by ATR-FTIR and the spectra obtained was more than 22 spectra as they are also replicates one. All digitalized spectra data were obtained for further analysis with MATLAB software. For ethanol standard samples, the preparation and analyzation was conducted the same day. Figure 3.1 below are ATR-FTIR spectra collected for ethanol samples:

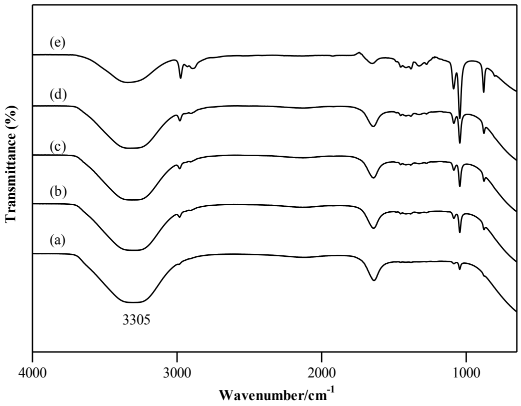

Figure 3.1 ATR-FTIR spectra of ethanol with concentration of (a) 10 % v/v, (b) 30 % v/v, (c) 50 % v/v, (d) 70 % v/v and (e) 90 % v/v

Analysis for ethanoic acid standard samples were conducted on a different day than ethanol standard samples. Figure 3.2 below are ATR-FTIR spectra for ethanoic acid standard samples.

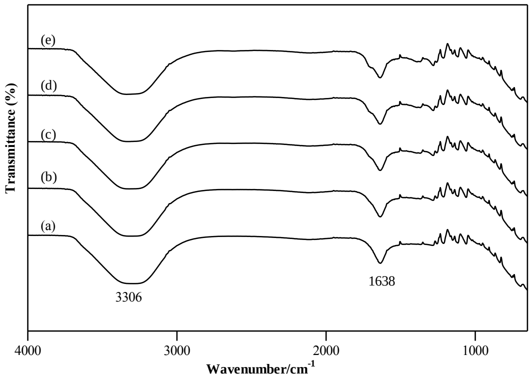

Figure 3.2 ATR-FTIR spectra of ethanoic acid with concentration of (a) 10 % v/v, (b) 30 % v/v, (c) 50 % v/v, (d) 70 % v/v and (e) 90 % v/v

From all spectra of ethanol and ethanoic acid standard samples obtained, it shows no big difference in intensity between each spectrum. This is due to the differences in concentration as the concentration of both ethanol and ethanoic acid increases, the intensity of the peak decreases especially in O-H band area.

Table 3.1 Infrared band assignment for all commercialized vinegar

Frequency (cm^{-1})	Intensity	Functional group and mode of vibration
3308.41	Strong	O-H stretching
2951.03	Weak	C-H stretching
1637.66	Strong – Moderate	C=O stretching
1245.59	Weak	C-O stretching

Most of the wavelength of the spectra of all commercialized samples of all the bands listed in Table 3.1 above were around the same value as shown in the figures below.

3.2. Conversion of Data

As shown in the IR spectra figures above, the data obtained was in the form of % Transmittance (% T). However, in quantitative analysis, it is more preferred if absorbance was used instead of % T. This is because % T is much more suitable for qualitative analysis. Since this study is about quantitative analysis, it is wise to choose absorbance than % T. Therefore, conversion of data from % T into absorbance was done by using the following equation [5].

$$A = 2 - \log_{10} \% T$$

These steps must be taken before importing all data into MATLAB and PLS_Toolbox for further analysis.

Table 3.2 Converting transmittance into absorbance

Absorbance (optical density)	Light Transmittance (%)
0	100
1	10
2	1
3	0.1
4	0.01
5	0.001
6	0.0001

According to the table above, at an absorbance of 6, only 10,000th of one percent of a particular wavelength is being transmitted through the filter. Absorbance is measured with a spectrophotometer, which establishes the light transmission and calculates the absorbance. However, the spectrophotometer can only measure absorbance up to 4.5 directly. Beyond this level all values must be extrapolated. For example, if a 2mm thick filter is measured to have an absorbance of 3, then it is assumed that a 4mm thick filter should have an absorbance of 6 [5].

3.3. Principal Component Analysis

In order to perform the Principal Component Analysis (PCA), the data matrix 42×1676 samples was imported into PLS_Toolbox to transform the original variables into principal components. Before the actual transformation, the data was pre-processed using mean centre method. This method was chosen because all variables that are used are in the same unit which is absorbance. If there are any differences in unit, other pre-processing mode should be used [6].

PCA was carried out first to find the region or section of the data that gives the most impact. As shown in Figure 4.5, data from the range of 1000 to 1600 gives the most impact specifically in the 1400 – 1500 region. From this result, data in the range of 1000 to 1600 was selected during analysis of Partial Least Squares (PLS) to build a PLS Regression model and testing out the model using data of commercialized vinegars.

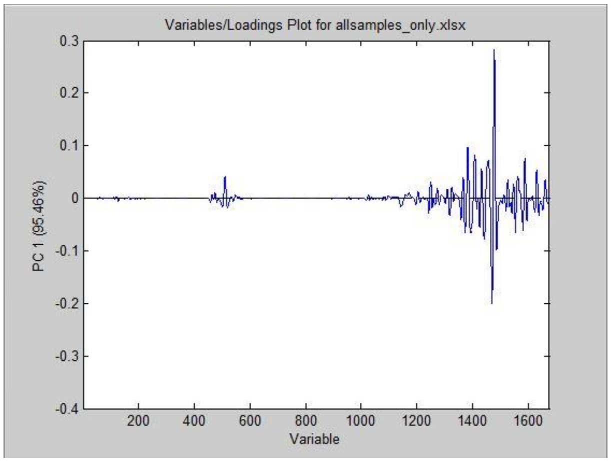

Figure 3.3 Loadings values of Principal Component Analysis of all samples

3.4. Principal Least Squares

The data imported into X-block and Y-block was in the form of matrix as MATLAB can only read data in the form of matrix. From PCA, a range of data was selected which is between 1000 to 1600 as shown in the figure below. This is because, data that gives the most impact is within the range of 1000 to 1600.

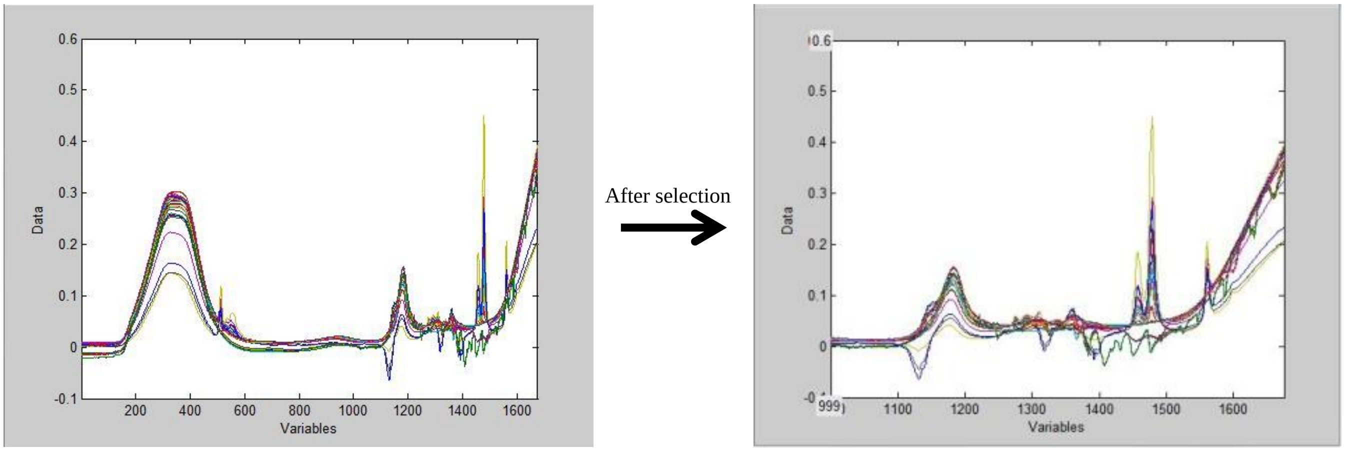

Figure 3.4 Before and after data selection

The pre-processing of X-block data used was the same as PCA because as stated before, since the unit used was all the same mean center mode was used. Cross-validation has to be done because cross-validation calculates the predictive ability of potential models to help in determining the appropriate number of components to retain in the model. Cross-validation is best if the optimal number of components was known [7]. In PLS regression, the cross-validated fitted value is the predicted response for each observation in data set, calculated individually, so the observation can be excluded from the model used to calculate the predicted response for that observation.

Cross-validation has to be done because cross-validation calculates the predictive ability of potential models to help in determining the appropriate number of components to retain in the model. Cross-validation is best if the optimal number of components was known [7]. In PLS regression, the cross-validated fitted value is the predicted response for each observation in data set, calculated individually, so the observation can be excluded from the model used to calculate the predicted response for that observation.

After all data has been processed and built into a model, four latent vectors (LV) or also known as PLS factors were obtained from the model. These LV are a special set of vectors associated with a linear system of equations (i.e., a matric equation) [7]. Based on Figure 4.10, it can be proven that four LV are the best one as suggested by MATLAB. By using robust component selection (RCS) statistic the optimal number of latent variables in the PLS regression model was selected [8].

$$\mathrm{RCS_k} = \sqrt{\gamma \mathrm{R} - \mathrm{RMSECV_k^2} + (1-\gamma)\mathrm{R} - \mathrm{RMSE_k^2}} \qquad (1)$$

With three different setting on the tuning parameter lambda (ɤ = 0.27, 0.87, 0.3), the RCS as a function of increasing number of LV showed a similar behavior with a decline and after four variables [8] as shown in Figure 3.5 below. Therefore, only four LV remained in the final predictor as suggested by the software. From the RMSEC (green) as shown in Figure 3.5, lambda = 0.3 indicates that only the good-of-fit remains in the RCS. Meanwhile, for RMSECV (blue) where lambda = 0.87 means that only the contribution of the quality of predictions remains in the RCS though it is better if the lambda = 0 for RMSEC and lambda = 1 for RMSECV [8].

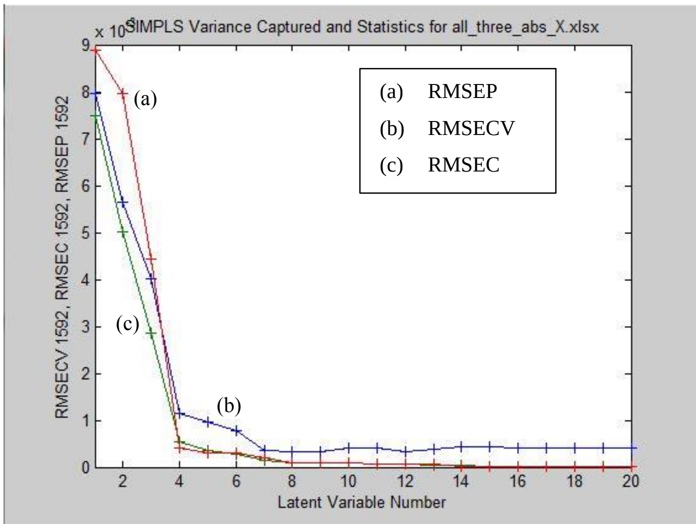

Figure 3.5 Graph of Root Mean Squared Error of Calibration (RMSEC) against number of Latent Variables where (a) RMSEP (ɤ = 0.27), (b) RMSECV (ɤ = 0.87) and (c) RMSEC (ɤ = 0.3)

In Figure 3.5, the standardized residuals of Y against Scores on LV 4 was plotted and ethanol with 90 % v/v concentration and its replicates (90a, 90b, 90c), ethanol with 70%v/v concentration (70c), and first replicates of sample A, B, C and D (A1,B1,C1,D1) are outliers as it stray off from the group. However, it all had a positive influence on the concentration of ethanol and ethanoic acid quantitative model, thus, it was not removed from the dataset. No outliers were discarded as the result in their exclusion will cause poor extrapolation of the model [8].

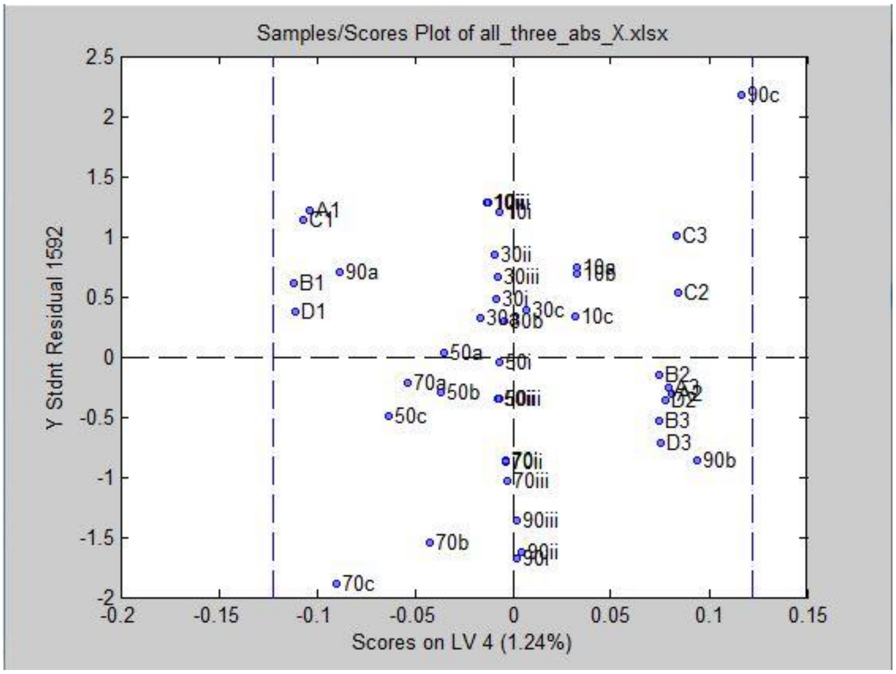

Figure 3.6 Distribution of all samples by standardized residual and scores on LV4

A PLS regression model was generated and validated by cross-validation. To evaluate the predictive ability of the model and to determine the optimum number of PLS factors or latent variables which in this case is 4, the Root Mean Square Error of Prediction (RMSEP) was calculated from these cross validation data as:

$$\mathrm{RMSEP} = \sqrt{\frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}}(\hat{\mathrm{y}}_{\mathrm{i}} - \mathrm{y}_{\mathrm{i}})^2} \qquad (2)$$

where n is the size of the test which is 42, $\hat{y}_i$ and y_i are respectively predicted and reference values of ethanol and ethanoic concentration (% v/v) in sample i (vinegar) [9].

From figure 3.7, a very good R^2 was managed to be obtained which is close to 1. Not only R^2 value obtained was good but all errors which are Root Mean Squared Error of Calibration (RMSEC) and Root Mean Squared Error of Cross-Validation (RMSECV) were also very satisfying. Even after data of real samples were tested using the model obtained, the value of R^2 still maintained very high and the Root Mean Squared Error of Prediction is very low. Compared to when using data from outside region 1000-1600, the value of R^2 change drastically from 0.982 to 0.031. This shows that selection of data is very crucial when building a PLS regression model and data at Y-column 1592 which is 818 cm^{-1} gives the best result when building this model.

From Figure 3.7, it can be seen that all data of real samples touched the regression line. This means that the all commercialized vinegars samples were within the range of standard samples which are from 10 % v/v to 90 % v/v of both ethanol and ethanoic acid. Since the correlation coefficient or R^2 value is great, 0.996 (calibration) and 0.999 (prediction), and all three errors, RMSEC, RMSECV and RMSEP have small values, 0.000535, 0.0011474, and 0.00010918 respectively, this model can be used for quantitative analysis of ethanol and ethanoic acid in vinegar.

RMSECV is as an internal indicator of the predictive ability of the model while RMSEP acts as a statistical measure for external cross validation. By using RMSEP, it can also expresses the average error to be expected on future prediction when calibration model is applied to unknown samples. Meanwhile, the smaller value of RMSECV indicates a better prediction ability of the model [10].

The correlation coefficient and RMSEP obtained from Table 3.3 and Figure 3.7 is 0.999 and 0.00010918 % v/v. Therefore, by using the equation for (2), the concentration (% v/v) of samples A, B, C and D can be calculated.

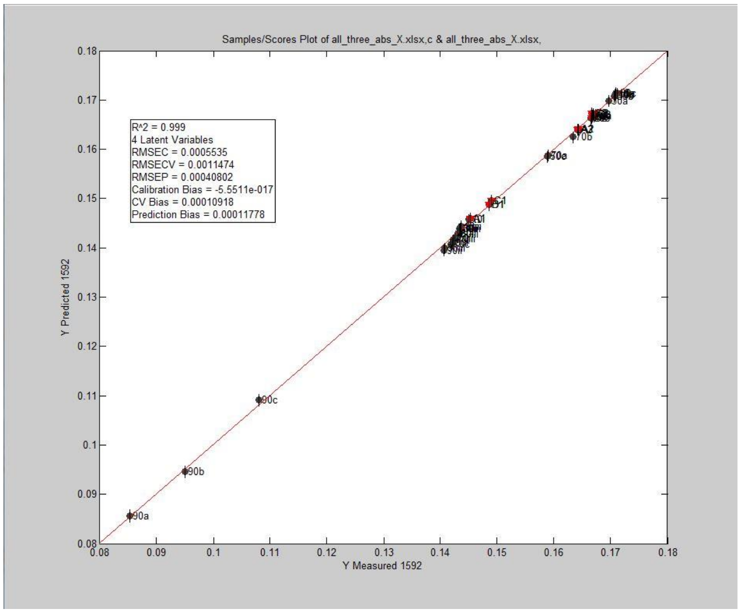

Figure 3.7 Regression line of PLS model after data of real samples were tested

Table 3.3 PLS model output from regression before(calibration) and after(prediction) data of real samples were tested

	Calibration	Prediction
Latent Variable	4	4
R^2	0.996	0.999
RMSEC	0.0005535	0.0005535
RMSECV	0.0011474	0.0011474
RMSEP	-	0.00040802
Calibration Bias	-5.5511e-017	-5.5511e-017
CV Bias	0.00010918	0.00010918
Prediction Bias	-	0.00011778

4. CONCLUSION

Partial Least Squares (PLS) regression is widely used to solve the chemical process problems. It is because the advance in computering technology makes things easier and it is also easily available. By using PLS regression, a chemist will be able to use more data and extract more info from spectra. In this study, ethanol and ethanoic acid in vinegar can be analysed using ATR-FTIR technique and a model was managed to be built using data gained from FTIR spectra. The model was applied to four different types and brands of commercialized vinegar. The data of real samples were tested and the results obtained was satisfying as all errors, SMREC, SMRECV and SMREP have small values which is 0.000535, 0.0011474, and 0.00010918 respectively also the value R^2 is close to 1.

REFERENCES

[1] San Chiang Tan, Vinegar Fermentation. Master of Science (Food Science) Thesis Report. (2005). Louisiana State University.

[2] M. Cirlini, A Caligiani, L. Palla, G. Palla, HS-SPME/GC-MS and chemometrics for the classification of Balsamic Vinegars of Modena of different maturation and egeing. *Food Chmistry*. **124** (2011), 1678-1683.

[3] Malaysian Food Regulations 1985. Section 34 of the Food Act 1983.

[4] Gunnar Gran, Determination of the Equivalent Point in Potentiometric Titrations. *Acta Chemica Scandinavica*. (1950) 559-577.

[5] IUPAC, *Compendium of Chemical Terminology*, 2nd ed. (the "Gold Book") (1997). Online corrected version: (2006–) "Absorbance”.

[6] Mustapa Nurul Syifaa', Detection of Honey Adulteration Using Fourier Transform Infrared Spectroscopy and Chemometrics. Bachelor of Science (Industrial Chemistry) Thesis Report. (2015). Universiti Teknology Malaysia.

[7] Usman Bishir, Prediction of Corrosion Inhibition Efficiency of Thiophene Derivatives Using Quantitative Structure Activity Relationship Method. Doctor of Philosophy (Chemistry) Thesis Report. (2015). Universiti Teknology Malaysia.

[8] Lin, Z., et al. (2016). Evaluation of the Bitterness of Traditional Chinese Medicines using an E-Tongue Coupled with a Robust Partial Least Squares Regression Method. Sensors (Basel). **16**(2),151.

[9] Oleszko, A., et al. (2017). Comparison of FTIR-ATR and Raman spectroscopy in determination of VLDL triglycerides in blood serum with PLS regression. Spectrochim Acta A Mol Biomol Spectroscopy. **183**, 239-246.

[10] Helland K., Berntsen H.E., et al. (1991). Recursive algorithm for partial least squares regression. Chemometrics and Intelligent Laboratory Systems. **14**, 129-137