A statistical analysis of total soluble solid in pineapple fruits

Wiyada Tanvatanagul∗
Peena Saleewong

Department of Statistics
Faculty of Science
Chiang Mai University
Chiang Mai 50200
Thailand

Abstract

The objective of this research is to propose a model for predicting the sugar content or Total Soluble Solid (TSS) in pineapple flesh by measurements of light absorption at the wavelength in NIR stage of which fruits are not destroyed. In addition, relation between the sugar content in pineapple flesh that was predicted by multiple linear regression which derived from NIR data and sugar content obtained from chemical process is investigated. Moreover, we can apply our prediction method in screening the qualities of pineapple through the use of the reflectance \( R_\lambda \) data which is transformed to the quantity of absorption at the wavelength \( \lambda \) \( A_\lambda \). This research is conducted using 120 sampling units. There were 60 randomly assigned by SPSS/FW program for chemical process, and the rest for comparing sugar content obtained from this assumption with the prediction results using multiple linear regression and equation from Principal Component Regression (PCR).

Keywords : NIR, multiple regression, principal component analysis, correlation analysis.

1. Introduction

The pineapple measurements of total soluble solid (°Brix) can be tested by biochemistry by many processes and need a lot of fruits for testing such as squeezed juices for the chemical process, dried method, cut flesh method. These biochemical processes are expensive and require

∗E-mail: wiyada@chiangmai.ac.th

Journal of Interdisciplinary Mathematics
Vol. 9 (2006), No. 3, pp. 431–444
© Taru Publications
experts and yet not all the fruits can be measured even with large amount of time. Therefore, we use another indirect method for evaluation the quality of pineapple. The NIRS (Near Infrared Spectroscopy) is the technology for testing quality inside the fruit by using nondestructive measurements. Statistics methods are needed to search for the relation between reflection or the absorption of electricity magnet wave, which in the NIRS (the wavelength of 800-2,500 nanometers (nm)) and the quantity of substances which indicated the quality of fruits, such as the sugar content or total soluble solid (TSS) in the fruits, the flour in the fruit etc. To search for the best wavelength stage in indicating the quality of all fruits, we use statistical techniques such as prediction, the principal component analysis and the multiple regression analysis.

The objective of this study is

1. To create the equation which can calculate the total soluble solid in the flesh of pineapples by measuring from the quantity of absorption the electricity magnet wave which the wavelength stage of the NIR by nondestructive measurement of pineapples.

2. To find the relation between the total soluble solid in pineapples from the predicting equation NIR spectroscopy and the sugar content or quantity of sugar in pineapples flesh from the chemical process.

2. Basic terminology

For ease of reading, we list here the basic terminology.

Quality means the level of excellent goods which consists of the external characteristics, quality of flesh, taste, nutritious value and safety. This research assigned the total soluble solid or the sugar content (sweetness) as the quality index of pineapple.

Total quality means the quantity of all solids which can dissolve in the pineapple flesh.

Calibration means the equation which derived from calculating the data of NIR spectroscopy and the equation of the total soluble solid. The result equation is called calibration equation.

Validation means taking data from the NIR spectroscopy, which is the rest of calculated equation data, to test the calibration equation by using these data predict the total soluble solid (TSS) by calibration equation.
NIR Spectroscopy Method is the method to measure the ray absorption in the range of near infrared ray (NIR), which cause the vibration of molecules. The data, which got from the NIR absorption at any wavelengths can describe the particular characteristic of the substance. In the same time, it will also tell the quantity of the substance. This NIR technique can be applied for checking food quality by nondestructive measurement to the food.

The important characteristics of NIR Spectroscopy are listed as follows.

• There is no need to use lots of chemicals as in standard general chemical analysis. This resulted in economical profits as well as no pollution in the laboratory.
• Be able to analyze rapidly with easily prepared sample to test.
• No need technicians who are experts in the analysis. Once the analysis is complete, it can be ready used.
• Only one sample can make many tests.
• Be able to know many types of information from only one test.
• Quick analysis and also suitable for the quality controlling which is used in factories.
• Be able to measure all products by adapting the checker in any places and on-line at any time.
• Be able to use in the vibrant or the movement place such as on a ship, because it’s not relied on weighing.

3. Data analysis

The data that are used in this research come from Department of Biology, Faculty of Science, Chiang Mai University. The data were as follows.

1. The values of reflection at the wavelength 400-1,700 nanometers \( (R_{\lambda}) \) from the NIR spectroscopy got from 120 pieces of pineapple flesh laying down in the NIR spectroscopy instrument Perten Diode Array DA-7000. The instrument shot the spectrum ray on the pineapple pieces so as to search for the concentrated of the ray of reflection in every wavelength of 5 nanometers. The value of reflection of each sample ray is the average from 3 times of shooting spectrum ray. Change the value result of reflection to the absorption by using the Beer-Lambert rule.
\[ A_\lambda = \log(1/R_\lambda) \]

\[ A_\lambda \] is the quantity of absorption at the wavelength \( \lambda \).

\[ R_\lambda \] is the quantity of ray reflection at the wavelength \( \lambda \).

2. The quantity of substances which indicated pineapples’ quality by chemical processing derived from the pineapple which passed the reflexive examination to test for the total soluble solids (TSS) by the chemical process.

To analyze data with SPSS/FW program, we used statistics for analysis as follow:

(1) The descriptive statistics explain the factor characteristics by the average and standard deviation.

(2) The statistical inference techniques consist of

- The multiple linear regression analysis to predict the equation of the total soluble solid in pineapple flesh with the value of absorption at the suitable wavelength stage.

- The principal component analysis to search for the wavelength stage of substances which indicate the total soluble solid in pineapple flesh (TSS).

- The correlation coefficient analysis to find the relation between the total soluble solid from the equation and from the chemical process, then test relation between them.

Step of data analysis

1. Divide the data into 2 groups by using SPSS/FW program simply sampling and assigned the sample size to be 50% of the whole data set.

2. For the first group of data, it will be measured by the calibration equation which can predict total soluble solid from the value of absorption and estimate the rest data (the second group) to measure TSS by chemical process.

3. Find the relation between the total soluble solid from the prediction and the total soluble solid from the chemical process.
4. Results

4.1 The general data of total soluble solids and the value of absorption at wavelengths

From 60 pieces of pineapple flesh, we found that the minimum value of TSS is $7.60 \degree$ Brix, maximum is $18.80 \degree$ Brix, average is $12.90 \degree$ Brix. The standard deviation is $2.58 \degree$ Brix. The average of absorption at any wavelengths is closed to each other and gradually decreased when the wavelength value increase, for the wavelength at 815 nanometers. The average of absorption is $-7.94$ and when the wavelength value ups to be at 1,000 nanometers, the average of absorption value is $-8.27$. The standard deviation of the absorption will be decreased when the wavelength increased. The spread of absorption at the high wavelength will be less than the low absorption.

4.2 Calibration equation of the total soluble solids in any stages by the multiple regression analysis

If $K$ is independent variables $(X_1, X_2, \ldots, X_k)$ which related to the dependent variable $Y$ by the linear combination, it will get the multiple regression equation which showed the relation between $Y$ and $X_1, X_2, \ldots, X_k$ as:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + e.$$ 

Let

- $\beta_0$: Line intercept the vertical axis (Y-intercept) or constant at $X_1 = X_2 = \ldots = X_k = 0$;
- $\beta_1, \beta_2, \ldots, \beta_k$, are the regression coefficient or $\beta_i$; $i = 1, 2, \ldots, k$
- $\beta_i$ are the ratio between the vertical change and the horizontal change along the line;
- $e$: Random error;
- $R^2$: the measure of how well a regression model fits. This statistic represent how much of the variance in the response is explained by weighted combination of predictors. The closer $R^2$ is to 1, the better the model fits\[14]\.

For this research,

- Independent variables (predictors) are the electricity magnet absorption of the pineapples which stage 800, 805, 810, 815, \ldots, 1700 nanometers.
Dependent variable (response) is the quantity of substances which indicates the pineapple quality (TSS).

4.2.1 Multiple linear regression in term of absorption

At the level of significance 0.01, prediction of the total soluble solids (TSS) from any wavelength (the independent variables) by stepwise method and weighted estimation in regression analysis. We found that there is 5 values of absorption wavelength 950, 1285, 1365, 1415 and 1665 nanometers. For this equation, the absorption value at that wavelength can explain the variation of total soluble solids \( R^2 \) 0.930 or 93.0% and standard error of estimation as 2.50065 by the equation as:

\[
TSS = -20.850 + 81.584(A_{950}) - 257.868(A_{1285}) + 421.185(A_{1365}) - 36.374(A_{1415}) - 211.507(A_{1665}).
\]  

(A)

From equation (A), we get the equation in the pattern of reflection as follow:

\[
TSS = -20.850 + 81.584(\log_{10} R_{950}) - 257.868(\log_{10} R_{1285}) + 421.185(\log_{10} R_{1365}) - 36.374(\log_{10} R_{1415}) - 211.507(\log_{10} R_{1665}).
\]  

(B)

Where as \( R_\lambda \) is the quantity of ray reflection at the wavelength \( \lambda \).

4.2.2 Multiple linear regression equation in term of the differences between \( A_\lambda \) and mean \( (A_\lambda) \)

The cause of using the difference or deviation is from the residual variation in equation (A) are not constant. Therefore we will decrease residual variation by using the deviation of data from their means to analyze the model.

Where as \( A_\lambda \) is the quantity of absorption at the wavelength \( \lambda \).

Mean \( (A_\lambda) \) is the mean of \( A_\lambda \).

d\( (A_\lambda) \) is the differences between \( A_\lambda \) and mean \( (A_\lambda) \) at the wavelength \( \lambda \).

At the level of significance 0.01, prediction of the total soluble solids (TSS) from any wavelength (the independent variables), found that there is 4 values of absorption wavelength 800, 820, 935, and 1075 nanometers. For this equation, the absorption value at that wavelength can explain
the variation of total soluble solids ($R^2$) 0.856 or 85.60% and standard error of estimation as 1.7737 by the equation as:

$$TSS = 12.752 - 68.348d(A_{800}) + 57.985d(A_{820}) - 25.282d(A_{935}) + 37.988d(A_{1075}).$$  \hspace{1cm} (C)

From equation (C), we get the equation in the pattern of reflection as follows:

$$TSS = 12.752 - 68.348d(\log_1/R_{800}) + 57.985d(\log_1/R_{820}) - 25.282d(\log_1/R_{935}) + 37.988d(\log_1/R_{1075}).$$  \hspace{1cm} (D)

For the testing assumption of multiple regression equation, at the level of significance 0.01, we accept all of these assumptions.

4.3 Calibration equation of the total soluble solids by the Principal Component Regression (PCR)

From multiple linear regression by stepwise method in topic 4.2. The multicollinearity between independents variables in equation will be removed from equation. Some these independents variables which are high related to dependent variable can not be in this equation. There are many methods of quantitative analysis statistics about NIR Spectroscopy. For this research, we used the Principal Component Regression (PCR) [3].

This method combines the Principal Component and the Inverse Least Square Regression together to solve the calibration equation. Therefore, the process of Principal Component Regression is divided into 2 steps:

1. Principal Component Analysis.
2. Regression Analysis.

4.3.1 Principal Component Analysis (PCA)

PCA is one of the extraction factor technique in many methods of factor analysis. Decreasing variable is the objective of PCA by using the relation of linear combination of the observed data, which no assumption about the relation of causally and reasonably between the factors and the variables. In the other hand, PCA is used for decreasing the complicated of variables which may affect to many factors.
4.3.2 Regression Analysis

Only the output of PCA can not predict the concentration of substances. We must use the multiple regression analysis technique from Inverse least Square model in the prediction or total soluble solids.

We chose one of interesting substances such as substance C in the base of the ray absorption at many wavelengths in the constant in difference absorption. From the hypothesis that the ray absorption in each wavelength has the linear combination with the substance C in the protein, it will be the following equation:

Let $A_{ij}$ be the absorption of the sample $i$ and wavelength $j$.

$P_i$ is coefficient value of $A_{ij}$ at the wavelength $j$

$$P_1 A_{11} + P_2 A_{12} + P_3 A_{13} = C$$

Suppose at the wavelength $A_{ij}$, 3 the chosen value are $A_{11}$, $A_{12}$, and $A_{13}$. The constant $P_i$ is from Beer-Lambert’s rule.

$$A_i = \epsilon_i I_i C_i$$

$$\left(1/\epsilon_i I_i\right) A_i = C_i$$

Therefore

$$P_i = 1/\epsilon_i I_i$$

For $n$ standard samples, there will be the equation sets:

$$P_1 A_{11} + P_2 A_{12} + P_3 A_{13} = C_1$$
$$P_1 A_{21} + P_2 A_{22} + P_2 A_{23} = C_2$$
$$\vdots$$
$$P_1 A_{n1} + P_2 A_{n2} + P_2 A_{n3} = C_n$$

We use Multivariate Linear Least Square Regression to find the best value of $P_i$, for the unknown value of absorption $A^*_1$, $A^*_2$ and $A^*_3$ for the 3 values of wavelength and substance C in protein, it will have the equation as follow:

$$C^* = P_1 A^*_1 + P_2 A^*_2 + P_3 A^*_3$$

Similarly, we can transform the qualification of the samples such as the fat, the humidity or the flour, but we should aware in choosing of wavelength stage that suitable with each of sample qualification.
For this research, the score of factor $S$ from PCA is used to replace the absorption value $A_{ij}$ in the ILS model [3].

Therefore, model equation is as:

$$C = BS + E_{C}$$

Let

- $C$ is the matrix of substance concentrate which is its size $1 \times n$.
- $B$ is the matrix $1 \times m$ of regression coefficient.
- $S$ is the matrix of $m \times n$ of factor score from PCA.
- $E_{C}$ is the matrix $1 \times n$ of error.
- $n$ is the number of sample size (spectrum).
- $m$ is the common factor.
- Coefficient $B$ can be found by the Regression Analysis.

Relation between the total soluble solid (TSS) and absorption at the wavelength $\lambda$ by Categorical Principal Component Analysis: CATPCA\textsuperscript{[10]} as Figure 1.

![Component Loadings](image)

**Figure 1**

Relation between the total soluble solid (TSS) and absorption at the wavelength $\lambda$ ($A_{\lambda}$ or $AB_{\lambda}$)
From Figure 1, $A_B$ or $A_\lambda$ are the quantity of absorption at the wavelength $\lambda$; $\lambda = 800, 805, 810, 815, \ldots, 1700$. From the chosen wavelength (800 nanometers to 1,250 nanometers) which believe it’s the best stage for prediction the total soluble solid. When analyzing the total soluble solid at these wavelength by principal component analysis, we got 8 factors from extraction factors and 8 factors score (fac1, fac2, \ldots, fac8) by SPSS/FW program.

We chose the wavelength (800 nanometers to 1,250 nanometers) which was believed that it’s the best stage for prediction the total soluble solid. After that we analyzed the total soluble solid at this wavelength, then we have 8 factors from extraction factors and 8 factors score.

Predicting equation of the factors score are:

$$
\text{fac1} = 0.010Z_{800} + 0.010Z_{805} + 0.010Z_{810} + 0.011Z_{815} + \ldots + 0.011Z_{1250} \quad (1)
$$

$$
\text{fac2} = 0.074Z_{800} + 0.075Z_{805} + 0.074Z_{810} + 0.075Z_{815} + \ldots - 0.052Z_{1250} \quad (2)
$$

$$
\text{fac3} = 0.449Z_{800} + 0.432Z_{805} + 0.403Z_{810} + 0.379Z_{815} + \ldots + 0.154Z_{1250} \quad (3)
$$

$$
\text{fac4} = -1.839Z_{800} - 1.222Z_{805} - 1.208Z_{810} - 0.713Z_{815} + \ldots + 0.123Z_{1250} \quad (4)
$$

$$
\text{fac5} = -3.396Z_{800} - 2.764Z_{805} - 2.508Z_{810} - 1.903Z_{815} + \ldots + 1.134Z_{1250} \quad (5)
$$

$$
\text{fac6} = -0.005Z_{800} - 0.017Z_{805} + 0.430Z_{810} + 0.255Z_{815} + \ldots + 0.341Z_{1250} \quad (6)
$$

$$
\text{fac7} = 6.428Z_{800} + 4.825Z_{805} + 3.460Z_{810} + 2.207Z_{815} + \ldots + 3.737Z_{1250} \quad (7)
$$

$$
\text{fac8} = 10.727Z_{800} + 6.358Z_{805} + 5.656Z_{810} + 2.659Z_{815} + \ldots - 6.432Z_{1250} \quad (8)
$$

$Z_\lambda$ is standard score of $A_\lambda$ for $\lambda = 800, 805, \ldots, 1250$.

For the multiple regression analysis by stepwise method, at the level of significance 0.01, we found that the independent variables which are able to predict total solids are fac1, fac3, fac4, fac5, fac7 and fac8 getting the equation as follow:

$$
\text{TSS} = 12.844 + 4.28\text{fac1} - 1.14\text{fac3} + 1.047\text{fac4} + 0.524\text{fac5} - 0.535\text{fac7} - 0.742\text{fac8} \quad (9)
$$
By $R^2 = 0.998$, has the standard error in estimation = 1.00535. These factors score in equation can explain the variation of total soluble solids ($R^2$) 0.998 or 99.80%. For the testing assumptions of multiple regression equation, at the level of significance 0.01, we accept all of these assumptions.

4.4 Analysis of correlation between the total soluble solids from calibration equation and the total soluble solids from the chemical process

In this research, we used the combination test of the total soluble solid from the prediction and the total soluble solid from the chemical process. To see how total soluble solids (TSS) from prediction is almost different from total soluble solids from chemical process (Real TSS), we look at correlation coefficient and also correlation coefficient ($r$), standard error of prediction as Table 1 and Table 2.

Table 1

<table>
<thead>
<tr>
<th>Real TSS</th>
<th>Correlation ($r$)</th>
<th>Standard error</th>
</tr>
</thead>
<tbody>
<tr>
<td>TSS from multiple regression</td>
<td>0.925*</td>
<td>1.41550</td>
</tr>
</tbody>
</table>

*Significance at the level of significance 0.01

At level of significance 0.01, total soluble solids from prediction has the linear relation to the total soluble solids from real test and the correlation coefficient is 0.925. The total soluble solids prediction and the total soluble solids from chemical process are relation in the same direction at quite high. The standard error in this prediction is 1.41550.

Table 1

<table>
<thead>
<tr>
<th>Real TSS</th>
<th>Correlation ($r$)</th>
<th>Standard error</th>
</tr>
</thead>
<tbody>
<tr>
<td>TSS from PCR method</td>
<td>0.998*</td>
<td>1.5965</td>
</tr>
</tbody>
</table>

*Significance at the level of significance 0.01
From the equation (9), at the level of significance 0.01, we found that the total soluble solid which is predicted from equation and the total soluble solids from the chemical process are relation of each other, correlation coefficient is 0.998 and this relation is in the same direction highly. The standard error in this prediction is 1.5965

5. Conclusion and discussion

Our research reveals that the two prediction equations are capable of predicting the new sample group similarly but the equation from PCR is complicated and difficult to use. The equation from PCR has more error than the equation from multiple regression, because analyzing by PCR method in this research is chosen the unsuitable wavelength; the wavelength of 800-1,250 nanometers is chosen to use. It shows that some wavelength in this stage have no relationship to the total soluble solids, therefore the standard error of prediction is highly at 1.5965. As the equation from multiple regression choose the wavelength which most related with the totals solids by removing the wavelength that has no relation with the total soluble solids away by stepwise method. It made the total soluble solids quite more close to the real total soluble solids and understand the pattern of equation easily. If we know the value of ray reflection of samples (independent variables), it can predict the total soluble solids. As for the error, which may be occur by sampling, made the calibration samples become the unsuitable samples to estimate the validation. Therefore, it can be error or the total soluble solids test may be wrong by chemical process which affected the analysis result.

Suggestions for improvement of results.

1. The total soluble solid in flesh of pineapples calculated by our prediction method is closed to the total soluble solids from real test. The test process in measuring the sample absorption by the prediction equation and the sample for predicting should not be different in size, and quality of the pineapples in total soluble solids equation for prediction the total soluble solid in order to decrease the error in the test planning.

2. The amount of samples for this research is not enough. Larger sample size should improve our predicted results.

3. The period of limited time to collect data from the NIR spectroscopy instrument cause the error from the chemical process.
More time is needed collecting more efficient data for prediction the total soluble solids.

Advantages from this study

1. Our prediction equation allows to sort out the disqualify pineapples for many grades in order to quote the price in each grade.
2. Our prediction equation allows to sort out the quality pineapples rapidly without destruction them.
3. This analysis data steps for applications by using other fruits in order to find the index predict quality of those kinds of fruits.

References


*Received September, 2005*