CHAPTER IV

RESULTS AND DISCUSSIONS

This research work aims to classify three fragrant rice varieties (Kor Kho 15, KDML 105 and Pathum thani 1) using gas chromatographic profiles in conjunction with statistical methods. Four samples of the three rice varieties were collected from rice storage box every two weeks until twelve weeks. The 24 samples of each variety were taken and analysed by HS-GC. All 72 chemical profiles with 114 peak-area variables were used in this study. The statistical methods, such as LDA, SLDA were used to build the classification model.

4.1 Chemical Profiles in the Different Storage Durations

From the 114 peak-area variables in the 72 chemical profiles, the median and the box plot of peak-area variables in each rice variety are shown in Figure 4.1 and 4.2 as different storage duration.

Figure 4.1 shows the median value in each rice variety by a different storage duration from two weeks until twelve weeks. The median patterns in those durations were not different in the twelve week durations.

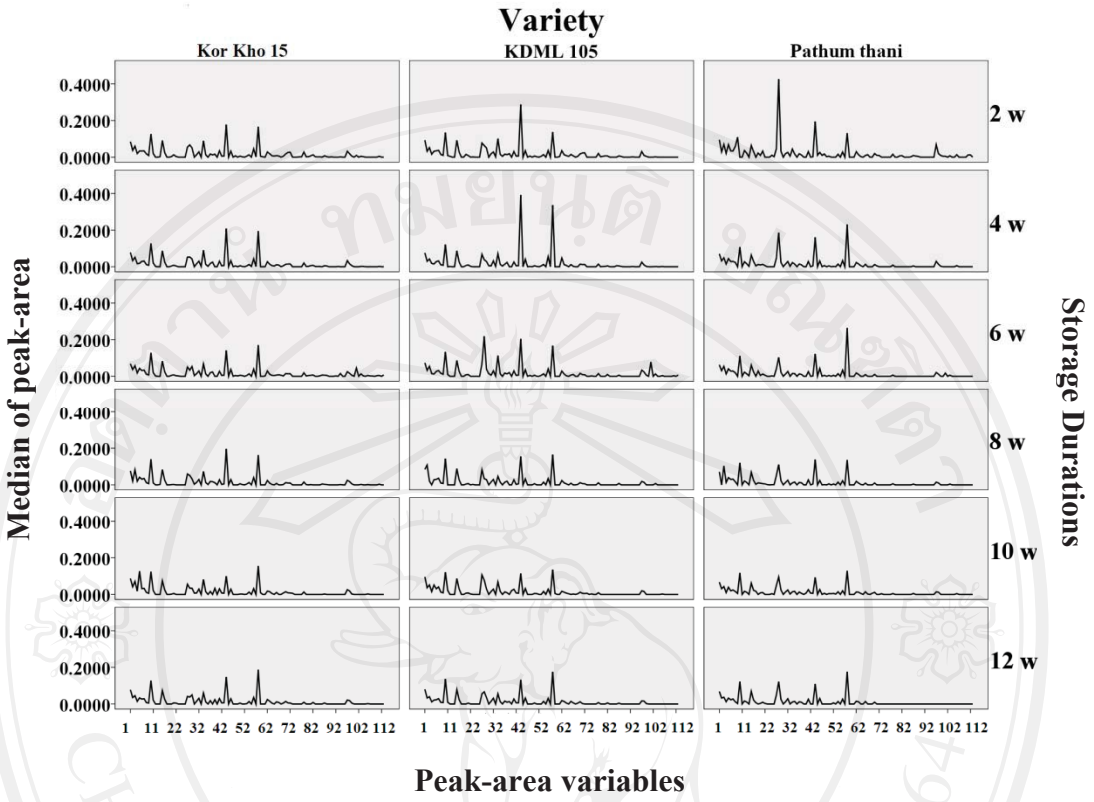

Figure 4.1 Median values of peak-area variables from different storage durations in each variety.

The box plot of peak-area values in the twelve weeks framework are shown in Figure 4.2 which shows that some of them are different for some duration, but most of peak-area value do not dramatically change over the twelve-week storage durations. The peak- area variations in the 6th week durations had higher variation in the profiles Kor Kho 15 and KDML 105. However, the peak-area numbers in these profiles are only a small part of the 114 peak-area variables. Therefore, the most of peak-area variation in each variety does not show a significantly different pattern.

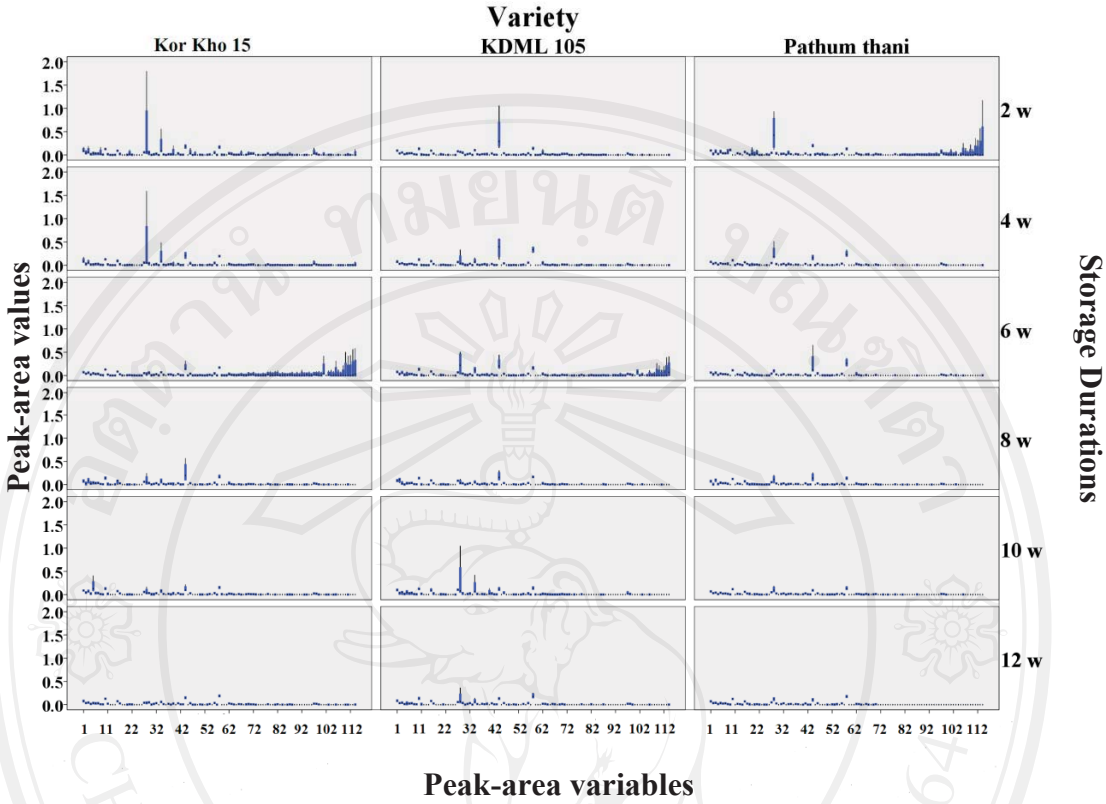

Figure 4.2 Box plot of peak-area variable in each storage duration and variety

In this study, we found that effect of storage durations (between 2 – 12 weeks) did not seem to change the quantity of peak-area values. However, the samples kept longer than twelve weeks may probably show a different pattern. In the twelve-week durations, the variation decreases because of the experimental-design package in this study. All of samples were kept in PE bag. The rice samples were controlled by the gas exchanged between rice sample and environments [14]. The samples were only exchanged between the samples in the same bag. For this reason, the peak-area variations in the longer storage durations are more closely related than the earlier durations [14]. For the peak-area pattern, it was found that the first-half of the profile of Figure 4.1 had median values with clear peak-area variables. Particularly, the first-

half of the peak-area variables were clearer than the second-half. From these peak-area patterns, only some parts of data which clearly express in their peak-area patterns were used to build the classification model.

4.2 Classification Results

Statistical methods, such as LDA and SLDA were applied to classify the fragrant rice varieties. Each sample had discriminant scores which were determined by the discriminant coefficient functions.

Since, the number of peak-area variables are higher than the sample sizes, LDA is not appropriate to analyse this data. Several methods were applied to reduce the number of variable. PCA is the method which could manage the 114 variables into the main components for classification model building, as shown in Table 4.1.

Table 4.1The main principal components from peak-area profiles

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	44.442	39.330	39.330	44.442	39.330	39.330
2	13.113	11.605	50.934	13.113	11.605	50.934
3	10.636	9.413	60.347	10.636	9.413	60.347
4	6.717	5.944	66.291	6.717	5.944	66.291
5	4.922	4.356	70.647	4.922	4.356	70.647
6	3.242	2.869	73.516	3.242	2.869	73.516
7	2.941	2.603	76.119	2.941	2.603	76.119
8	2.507	2.219	78.338	2.507	2.219	78.338
9	2.244	1.986	80.324	2.244	1.986	80.324
10	1.985	1.757	82.080	1.985	1.757	82.080
11	1.961	1.735	83.816	1.961	1.735	83.816
12	1.494	1.323	85.138	1.494	1.323	85.138
13	1.331	1.178	86.316	1.331	1.178	86.316
14	1.282	1.135	87.451	1.282	1.135	87.451
15	1.187	1.050	88.501	1.187	1.050	88.501
16	1.128	.999	89.500	1.128	.999	89.500
17	1.022	.905	90.404	1.022	.905	90.404
18	.901	.798	91.202			
19	.811	.717	91.919			
20	.773	.684	92.603			
.	.	.	.			
.	.	.	.			
.	.	.	.			
113	0.000	0.000	100.000			

The main 17 principal components that explained 90.40% of the total variance (Table 4.1) were formed from the 114 peak-area variables. Hence, only the first 17 principal components were enough to be retained in subsegment analysis and used to construct the classification model by using LDA.

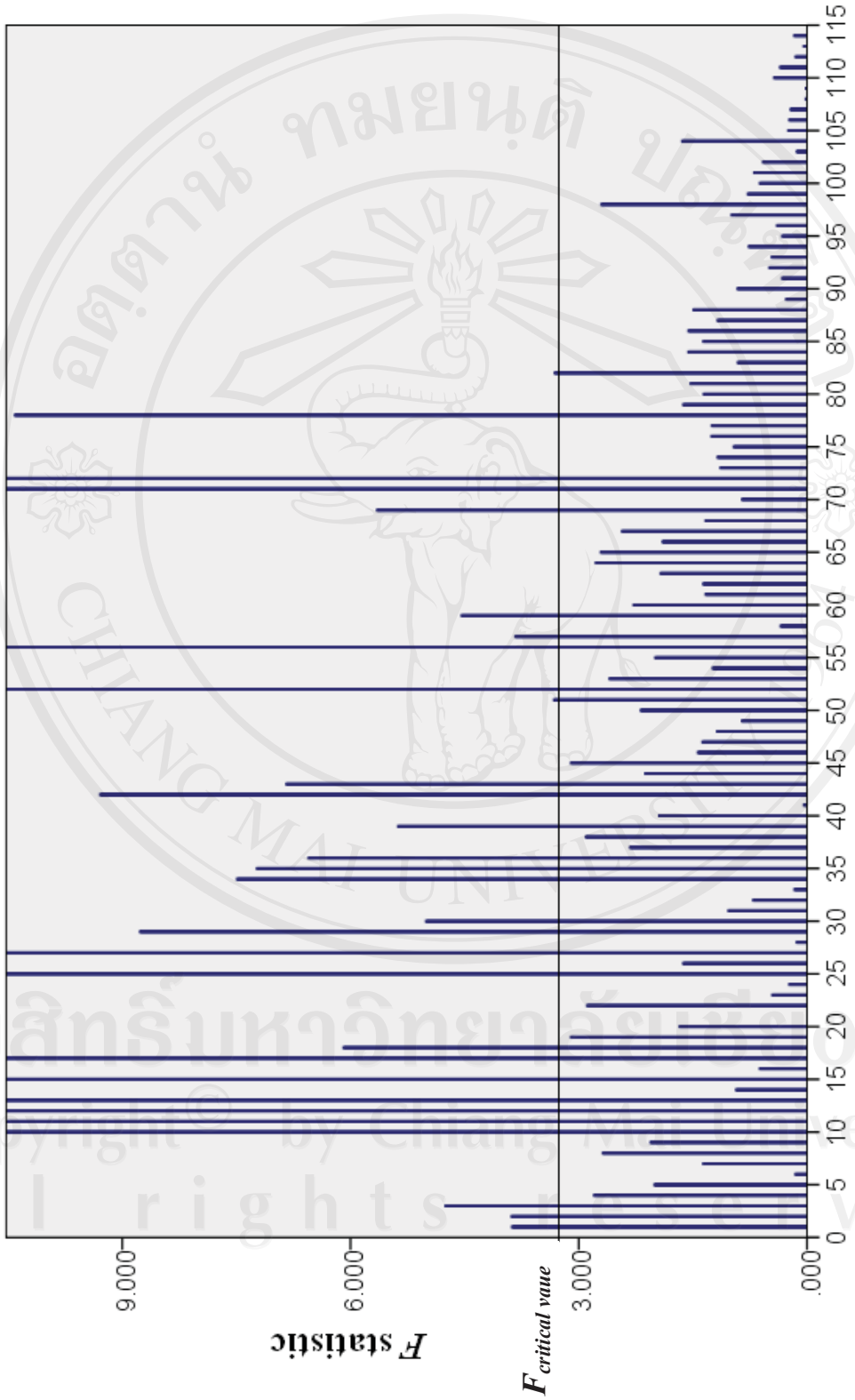

Figure 4.3 *F*-statistics from one-way ANOVA in each peak-area variables

The others method such as the selection variable method was used to reduced the number of variables. Firstly, the significant peak-area variables by using *F*-test in one-way ANOVA in each variety were evaluated. Figure 4.3 show the *F*-statistics of peak-area variables that received from the mean tested by one-way ANOVA at significant level 0.05. The significant peak-area variable is the variable that its *F*-statistic is higher than *F*-critical value (*F*-critical = 3.141). All 33 significant variables were used to construct the LDA and SLDA classification model.

The percentage of non-zero value samples in some varieties and percentage of non-zero value samples for all varieties in the different levels were selected the variable to construct the models. From the peak-area profiles, there are several peak-area variables of zero value. The peak-area variables of high zero-value frequency were removed and brought the rest variable to construct the classification models.

Some periods of the peak-area profile comparison were constructed the discrimination of the peak-area profile comparison were constructed the discrimination models. From the pattern of peak-area profiles in Figure 3.1-3.3, it shows the main peak-area variables which main variables were measured in the half first of profiles.

Every classification model was evaluated for the optimal model. The corrected classification and prediction results were presented as a "corrected classification (prediction) result" pattern. In this study, the leave one out-cross validation (LOOCV) technique was used to evaluate the classification models (Table 4.2).

Table 4.2 Classification results of the Thai-fragrant rice variety using the peak-area profile data in the different perspectives with discriminant analysis.

Variable Selection	% Classification Accuracy (Prediction) Using with			
	PCA with LDA	LDA	SLDA	LDA*
All peak-area variables	90.3(87.5)	-	100.0(98.6)	100.0(95.8)
Some period of peak-area profiles in the time range of (min)				
10-70	-	-	100.0(93.1)	98.6(94.4)
15-70	-	-	100.0(97.2)	98.6(94.4)
5-35	-	100.0(91.7)	97.2(97.2)	98.6(94.4)
10-35	-	100.0(90.3)	98.6(97.2)	98.6(97.2)
15-35	-	98.6(84.7)	94.4(93.1)	94.4(88.9)
5-25	-	98.6(90.3)	98.6(94.4)	94.4(87.5)
10-25	-	98.6(93.1)	98.6(95.8)	94.4(84.7)
15-25	-	91.7(87.5)	91.7(90.3)	90.3(86.1)
Percentage of non-zero value samples in some varieties				
≥ 50%	-	100.0(79.2)	95.8(93.1)	100.0(97.2)
≥ 75%	-	100.0(93.1)	98.6(97.2)	97.2(88.9)
≥ 95%	-	100.0(88.9)	100.0(98.6)	98.6(93.1)
Percentage of non-zero value samples for all varieties				
≥ 50%	-	100.0(88.9)	100.0(95.8)	98.6(88.9)
≥ 75%	-	100.0(88.9)	98.6(97.2)	98.6(88.9)
≥ 95%	-	98.6(94.4)	98.6(97.2)	93.1(87.5)

LDA* is the significant peak-area variable which were constructed by LDA model

Table 4.2 show the classification results for comparison between the discrimination results from the peak-area profile data in the different perspectives; the all of peak-area variables and the PCs from all peak-area variables, the significant peak-area variables (ANOVA), the percentage of non-zero value levels, and some periods of peak-area profiles.

The percentage of corrected classification in all peak-area variables with SLDA model and the significant peak-area variables with LDA model are 100% that are higher than the PCs with LDA model (90.3%). However, the percentage of corrected prediction by using SLDA with all peak-area variables is 98.6% that is higher than the significant peak-area variables with LDA model and PCs with LDA model.

When we constructed the classification models by using LDA, SLDA and significant variables with LDA based on the peak-area variables that appeared in a given period of time, we found that almost percentage of corrected classification and higher than 90%. The percentages of corrected prediction in SLDA model are also higher than 90% for all period of time. The optimal model is the model from all peak-area profiles using the SLDA model. Under the peak-area profile in the time range of 10-35 min, the percentage of corrected classification by SLDA model, and the significant peak-area variables with LDA model are 98.6 and 97.2, respectively. However, the percentage of corrected classification and prediction are 100 and 97.2%, respectively by using SLDA model with the peak-area profile period in time ranges of 15-70 min. Nevertheless, this model spent 70 min per sample to measure the rice profiles. The peak-area profiles in the time range of 10-35 min period had spent 35 min per sample for measurement of the profiles. From the peak-area measurement, the model from peak-area profile in the time range of 10-35 min with SLDA reduced the investigation time in each sample. It can be increased in double numbers of sample more than the previous chemical profile measurement. The statistics and discriminant functions are shown in Appendix C. On the selection of peak-area variables with a given percent of non-zero value; sample, the percentage of correct

classification and prediction are more than 85% for all models. These results suggest that the classification of Thai fragrant rice using HS-GC profiles in conjunction with statistical methods can provided good identification accuracy.

Based on the highest percentage of correct classification and prediction, we proposed to use all of peak-area variables with SLDA model to classify rice varieties. The discriminant functions of this model were generated and tested, as shown in Table 4.3 and 4.4.

Table 4.3 Wilks' lambda and eigenvalues of SLDA model from all peak-area variables

Test of Function(s)	Wilks' Lambda	Chi-square	df	p-value	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1 through 2	0.001	414.952	38	0.000	113.997	93.6	93.6	0.996
2	0.114	130.257	18	0.000	7.767	6.4	100.0	0.941

Table 4.3 shows the Wilks' lambda statistics from the all peak-area variables with SLDA. The Wilks' lambda is a statistic used in particular by discriminant analysis as a measure of the class centers separation. Wilks' lambda is a number between 0 and 1. A small (close to 0) value of wilks' lambda means that the groups are well separated. The chi-square statistics is used to find a statistical value is shown the p-value in the fifth column. If the p-value is lower than the significant level (α=0.05), this function is suitable for rice classification. The both discriminant function are appropriate for classification of fragrant rice varieties.

The eigenvalues represent the discriminating power of each discriminant function, as shown in Table 4.3. There is one eigenvalue for each discriminant function. The ratio of the eigenvalues indicates the relative discriminating power of the discriminant functions. The two discriminant functions show a decreasing order of the function importance. The first and the second functions explained the total variation percentage of 93.6 and 6.4%. The last columns show the canonical correlation values, which agrees with the relationship between the chemical profiles and rice varieties. The canonical correlations from the first and second functions are 0.996 and 0.941.

Table 4.4 The stepwise linear discriminant functions from all peak-area variables

Peak Number (X_i)	Retention Time	Function	
		1	2
2	6.050-6.301	6.802	22.683
9	9.760-9.920	122.609	120.762
10	10.150-10.360	89.649	151.009
11	10.380-10.600	137.431	309.215
13	12.300-12.500	-1,143	254.270
24	17.250-17.450	-775.284	288.166
27	19.700-19.900	-26.992	-168.071
32	21.675-21.825	26.689	-290.096
42	26.375-26.525	-55.761	-187.743
44	27.875-28.075	228.892	140.952
52	29.850-29.975	-608.957	341.982
58	33.030-33.330	13.717	3.367
59	33.700-33.880	77.973	312.356
74	45.990-46.050	342.252	-229.463
78	47.210-47.380	70.748	36.078
82	49.810-49.930	-123.995	-97.610
86	50.970-51.100	133.550	379.337
97	55.302-55.475	67.315	25.757
101	57.470-57.680	-18.838	-26.970
	(Constant)	-10.801	-7.960

The discriminant functions can be applied to predict the rice varieties by the coefficient in each peak-area variable (Table 4.4), as followed:

$$\widehat{Y_1} = -10.801 + 6.802X_2 + 122.609X_9 + 89.649X_{10} + 137.431X_{11} - 1{,}143X_{13} - 775.284X_{24} - 26.992X_{27} + 26.689X_{32} - 55.761X_{42} + \cdots . + 67.315X_{97} - 18.838X_{101}$$

$$\widehat{Y_2} = -7.960 + 22.683X_2 + 120.762X_9 + 151.009X_{10} + 309.215X_{11} + 254.270X_{13} + 288.166X_{24} - 168.071X_{27} - 290.096X_{32} - 187.743X_{42} + \cdots + 25.757X_{97} - 26.970X_{101}$$

The peak number in Table 4.4 is related to the retention time in the second column. This model was constructed from all of peak-area variables with SLDA method. The classification result is 100.0% and the prediction result is 98.6%. The misclassification case is Kor Kho 15, but the prediction is KDML 105. In the other models, the main-misclassification predictions occur between Kor Kho 15 and KDML 105.

From the results in this study, the peak-area profiles from HS-GC technique can be applied to classify the fragrant rice varieties. The peak-area profiles of the fragrant-rice varieties were used the statistical methods for classification the varieties by stepwise linear discriminant analysis. Especially for some periods of the peak-area profiles in the time range of 10-35 min with the stepwise method are reduced the HS-GC measured time. The several statistics and discriminant functions were shown in Appendix C. However, the peak-area profiles are possible to be changed because of several factors including to seasonal, cultivation area, water quantity, storage durations, and sunlight [14, 23, 24, 27-34].

The classification results show that the Pathum thani 1 patterns (discriminant scores) could clearly be different from the Kor Kho 15 and KDML 105 patterns [54]. The Kor Kho 15 and KDML 105 are close pattern, which always appear in every statistical model leading to the most misclassification between those two varieties. The classification results from several models are harmonious to the three rice properties. The Kor Kho 15 and KDML 105 have the same rice quality (slender rice varieties with low gelationsation temperature, soft-gel consistency and high 2AP quantity) and the Kor Kho 15 was developed from the KDML 105 [55]. It could be said that the Kor Kho 15 and KDML 105 are essentially the same as Thai Hom Mali or Jasmine rice [55]. The peak-area profile patterns in those two varieties are also more similar than that of the Pathum thani 1.