CHAPTER 3: RESEARCH METHODOLOGY

3.1 Introduction

This chapter covers the research flow and the sample of the study whereby further elaboration on the data collection methods and the main characteristics of the dataset used in this paper.

3.2 Conceptual Framework

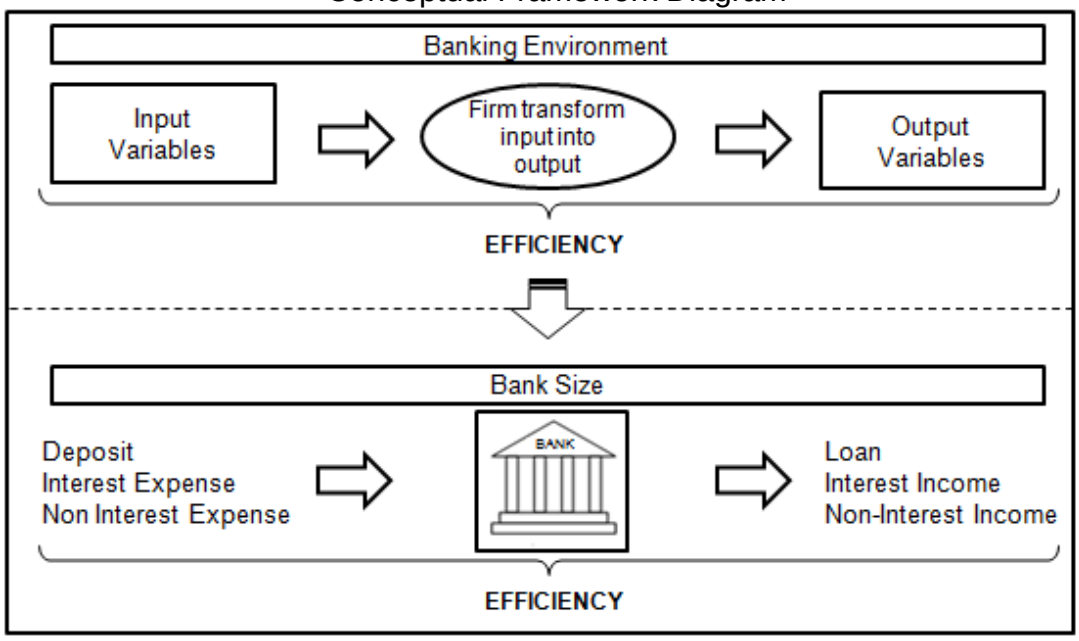
The conceptual framework of this study will be based on intermediation approach whereby banks are regarded as intermediary between savers and borrowers. The production approach is suitable for bank efficiency study (Berger & Humphrey, 1997).

Lozano et al. (2002) had made a study on commercial banks across 10 European countries investigating the operating efficiency and the environmental conditions using the general DEA model whereby the findings shows that environmental conditions exercise a strong positive influence over the behavior of each country's banking industry.

Figure 3.1 below illustrates the conceptual framework to determine if bank size matters in determining the bank's efficiency. The top portion of the diagram is adapted from Chu & Lim (1998) where the study on six Singapore listed banks was being evaluated relative to the cost and profit efficiency.

Based on the adopted framework, the same framework was being adopted in this study where the efficiency is being measured by the input and output variables and the impact of the banking environment that is contributing to it.

Figure 3.1
Conceptual Framework Diagram



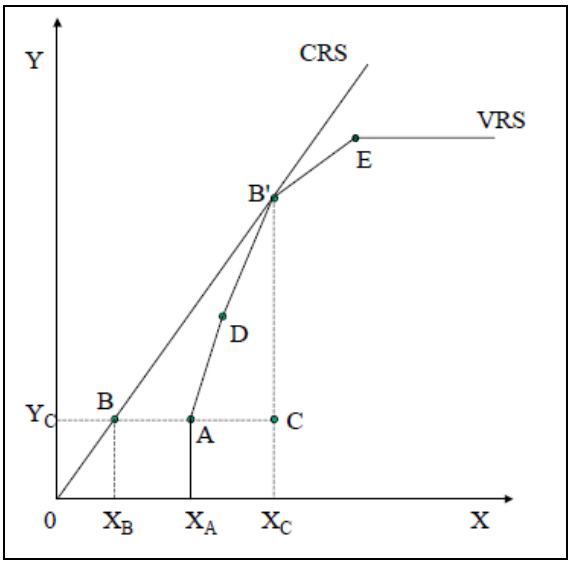
3.3 Selections of Measures

This study will applies a non parametric measure using Data Envelopment Analysis (DEA) by employing both constant return to scale (CRS) and variable return to scale (VRS) to compute the efficiency scores. The Malmquist Productivity Index technique is then employed to derive the index of productivity change. In the second stage, the Multiple Regression is used to identify potential influences of key bank specific environmental changes on the calculated bank efficiency measure.

The background of the DEA is a linear programming based technique to measure the relative efficiency of a fairly homogeneous set of decision making unit (DMU) multiple inputs to produce multiple outputs.

In production theory, the changes in output levels due to changes in input level is termed as return to scale where this can be further categorised as CRS and VRS. CRS implies that an increase in input levels by a certain proportion result in an increase in the output levels by the same proportion while VRS implies that an increase in the input levels however it need not necessarily result in a proportional increase in output levels. The linear relationship between the inputs and outputs is depicted in Figure 3.2 below.

Figure 3.2
Constant and Variable Return to Scale



The selection basis to use VRS opposed to CRS is justified on the basis that not all DMU are operating at optimal scale due to imperfect competition and financial constraints.

A bank operating in point C as in Figure 3.2 above is technically inefficient. The technical efficiency (TE) is measured by Y_cB/ Y_cC in an input orientation against the CRS frontier. The inefficiency can then be decomposed into pure technical inefficiency (PTE) and scale inefficiency (SE). The new measures

are calculated as PTE = Y_cA/ Y_cC and SE = Y_cB/ Y_cA. The TE can be defined as the product of pure technical efficiency and scale efficiency. The TE is how close the bank is to the production frontier and SE is how close the bank is to producing at optimal scale.

The DEA methamatical programming technique can be used to estimate the productivity improvement over time whereby it requires at least two time periods. The three measure of growth can be derived namely, total factor producitivy change, tehcnical efficiency change and technological change. The total factor productivity is derived by dividing an index of output production by an index of total input usage whereby it is represented by a ratio of all outputs produced in year t, (y_r^t) to an index of all inputs employed in ($f(x_r^t)$) in year t. (Grosskopf, 1993).

$$\text{TFP} = \frac{y_r^t}{f(x_i^t} = A(t) \qquad (3.30)$$

where:

y_r^t : all outputs *r* produced in year t

$f(x_r^t)$: all inputs *i* employed in year *t*

A(t) *:* fraction of $y_r^t / f(x_r^t)$

TFP is a combination of measure for the productivity of all inputs and outputs whereby TFP growth refers to the change in productivity over time is defined as the change in total factor productivity between two time periods, t and t+1 which can be formulated as,

$$\text{TFP change} = \frac{A(t+1)-A(t)}{A(t)} = \frac{y_r^{t+1}-y_r^t}{y_r^t} - \frac{f(x_i^{t+1})-f(x_i^t)}{f(x_i^t)} \quad (3.31)$$

where:

y_r^t : all outputs *r* produced in year t

$f(x_r^t)$: all inputs *i* employed in year *t*

A(t) : fraction of $y_r^t / f(x_r^t)$

The TE change is the ratio of the technical efficiency fro time period *t* to *t*+1.

$$\text{Technical Efficiency Change} = \frac{TE^{t+1}}{TE^t} \quad (3.32)$$

where:

TE^{t+1} : technical efficiency of year *t+1*

TE^t : technical effieicny of year t

The TE change measures if the firm is getting closer to the best practice frontier over time. This can be illustrated by firm learning from the best practice firms and improving the managerial practices of adopting a better system over time.

The Technology change is calculated from the ratio of the TFP change to the TE change as shown below:

$$\text{Technological change} = \frac{TFP\ change}{TE\ change} \quad (3.33)$$

Based on the measures above, Fare et al. (1994) have exteded this further which is also known as Malmquist Productivity Index (MPI). Based on Caves et al. (1982) this technique is the index of productivity change and do not require the cost or revenue to be aggregate input and outputs.

This study makes use of an input oriented (MPI) as it provides the best savings by cutting out the excessive use of inputs. By utilising Fare et al. (1994) model, the MPI can be formulated as follow:

$$M_0^t\left(X_{i,t}, Y_{i,t}, X_{i,t+1}, Y_{i,t+1}\right) = \left(\frac{d_0^t\left(X_{i,t}, Y_{i,t}\right)}{d_0^t\left(X_{i,t+1}, Y_{i,t+1}\right)} \times \frac{d_0^{t+1}\left(X_{i,t}, Y_{i,t}\right)}{d_0^{t+1}\left(X_{i,t+1}, Y_{i,t+1}\right)} \right)^{\frac{1}{2}} \quad (3.34)$$

where:

d_0^t : Distance function at time *t*

d_0^{t+1} : Distance function at time *t*+1

χ : Vector of inputs

Y : Vector of outputs

M_0 : Malmquist Productivity Index

The MPI represents the productivity of the production point $(X_{i,t}, Y_{i,t})$ relative to the production point $\left(X_{i,t+1}, Y_{i,t+1}\right)$. A value of greater than 1.0 indicates that there is an increase in the total factor productivity whereas the value below 1.0 represents a decline in total factor productivity.

Based on Fare et al. (1994), the MPI can be then decomposed into measures of technical efficiency change and technology change by factoring as shown in Formula (3.35) below.

$$M_0^t\left(X_{i,t}, Y_{i,t}, X_{i,t+1}, Y_{i,t+1}\right) = \left(\frac{d_0^t\left(X_{i,t}, Y_{i,t}\right)}{d_0^{t+1}\left(X_{i,t+1}, Y_{i,t+1}\right)}\right) * \left[\left(\frac{d_0{}^{t+1}\left(X_{i,t+1}, Y_{i,t+1}\right)}{d_0^t\left(X_{i,t+1}, Y_{i,t+1}\right)}\right) * \left(\frac{d_0{}^{t+1}\left(X_{i,t}, Y_{i,t}\right)}{d_0^t\left(X_{i,t}, Y_{i,t}\right)}\right)\right]^{\frac{1}{2}} \quad (3.35)$$

The first parenthesis is to measure the technical efficiency change which is represented by the relative distance from the input-output combination from the frontier in period *t* and *t+1*. Both the numerator and denominator of the ratio must be greater or equal to 1.0. If the technical efficiency is higher in period *t+1* than in period t, the value of this ratio will be greater than 1.0, vice versa. The second parenthesis represents the technology change between period *t* and *t+1*. Value greater than 1.0 imply technology progress.

The calculation of Malmquist Productivity Index exploit the fact that the input distance functions are reciprocal of Farrell's (1957) input-orientated technical efficiency measures whereby the DEA model can be used to calculate the distance functions with input orientation and CRS assumption. There will be four linear programs whereby the first two linear programs (Formula 3.36 and 3.37) are where technology and the observation to be evaluated are from the same period while the other two linear programs occur when reference technology is constructed from data in one period, the observation to be evaluated is from another period (Formula 3.38 and 3.39).

$$\left[D^{t}(y_t, x_t)\right]^{-1} = \min_{\lambda,\theta} \theta^{k'} \tag{3.36}$$

s.t. $$-y_{rj_t} + \sum_{j=1}^{n} \lambda_j y_{rj_t} \geq 0,$$

$$\theta_o x_{ij_t} - \sum_{j=1}^{n} \lambda_j x_{ij} \geq 0,$$

$$\lambda_j \geq 0$$

$$\left[D_I^{\,t+1}(y_{t+1}, x_{t+1})\right]^{-1} = \min_{\lambda,\theta} \theta \tag{3.37}$$

s.t. $$-y_{rj_{t+1}} + \sum_{j=1}^{n} \lambda_j y_{rj_{t+1}} \geq 0,$$

$$\theta\, x_{ij_{t+1}} - \sum_{j=1}^{n} \lambda_j x_{ij_{t+1}} \geq 0,$$

$$\lambda_j \geq 0$$

$$\left[D_I^{\,t+1}(y_t, x_t)\right]^{-1} = \min_{\lambda,\theta} \theta \tag{3.38}$$

s.t. $$-y_{rj_t} + \sum_{j=1}^{n} \lambda_j y_{rj_{t+1}} \geq 0,$$

$$\theta\, x_{ij_t} - \sum_{j=1}^{n} \lambda_j x_{ij_{t+1}} \geq 0,$$

$$\lambda_j \geq 0$$

$$\left[D_I^{\ t}(y_{t+1}, x_{t+1})\right]^{-1} = \min_{\lambda,\theta} \theta \tag{3.39}$$

s.t. $$-y_{rj_{t+1}} + \sum_{j=1}^{n} \lambda_j y_{rj_t} \geq 0,$$

$$\theta\, x_{ij_{t+1}} - \sum_{j=1}^{n} \lambda_j x_{ij_t} \geq 0,$$

$$\lambda_j \geq 0$$

Upon having the four linear program solved for each bank, and each pair of adjacent time period, the Malmquist Index can be calculated as its two components of efficency and frontier advances.

All DEA based efficiency and productivity estimation are conducted with the software EMS: Efficiency Measurement System version 1.3.

The Multiple Regression model will then be run to explain the environmental variables. Coelli et al. (1998) had suggested a few methods in which environmental variables can be accommodated in a DEA analysis. This research adopts the multiple regression model whereby based on the efficiency scores obtained from the first stage analysis are being regressed against the bank size dummy variable.

The standard Multiple Regression model (Formula 3.40) can be defined as follows for the observation.

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_k x_{ki} + \varepsilon_i \quad (3.40)$$

Where *Y* is the change that the programme is mainly suppose to produce which is the DEA score, x_i and β are vectors of explanatory variables and unknown parameters respectively while y_i^* is a latent variable and y_i is the DEA score. By using the efficiency scores as the dependent variable, we estimate the following regression model:

$$\begin{aligned}(TFPCH)_{jt} = \quad & \beta_0 + \beta_1(DEPO)_{jt} + \beta_2(INTEXP)_{jt} + \beta_3(NINTEXP)_{jt} \\ & + \beta_4(NLOAN)_{jt} + \beta_5(INTINC)_{jt} + \beta_6(NINTINC)_{jt} + \\ & \beta_7(SIZE) + \varepsilon_j \end{aligned} \quad (3.41)$$

where ‘j’ denotes the bank, ‘t’ the examined time period and ε is the disturbance term. The *(TFPCH)* is derived from the the CRS score, the *(DEPO)* captures the bank deposit, the *(INTEXP)* captures the interest expense, the *(NINTEXP)* captures the non-interest expense, the *(NLOAN)* captures the net loan, the *(INTINC)* captures the interest income, the *(NINTINC)* captures the non-interest income. *SIZE* is introduced in the regression model to examine the impact of bank size linked towards the bank’s total productivity levels.

The log linear form is used to improve the regression model's goodness of fit and may reduce simultaneity bias (De Bandt and Davis, 2000). However prior to performing the Multiple Regression, the variables are being tested for nomality.

3.4 Sample of the Study

The sample of this study is banks located in selected ASEAN countries whereby four different countries namely, Malaysia, Vietnam, Philippines and Thailand have been selected. The sample size takes into account the bank specialization whereby only banks categorized under commercial banks and specialized governmental credit institution are being selected. The bank specialization is based on BANKSCOPE definition.

3.5 Data Collection Methods

The purpose of this section is to describe the main characteristics of the dataset used and the steps used to build the dataset. The main source of the data is from BANKSCOPE which maintains a global database of banks' financial statements, ratings and intelligence. The components that have been extracted are the Balance Sheet, Income Statement and Total Asset.

The data is being downloaded individually for each bank based on the selected ASEAN countries to avoid duplication. The period of data obtained is from 1995 to 2009. To ensure the accuracy of the data obtained from

BANKSCOPE, random cross check against the data obtained and the data reported in the annual report is being performed.

There are instances data are not available for the some banks whereby the required data is being obtained from the bank's annual report. Banks with data less than 5 years are being excluded from this study. The final data set consist of 74 banks with a total of 370 bank years. Table 3.1 depicts the summary of banks that is being included in this research.

Table 3.1
Number of Banks Categorized by Country

Country	**Total**
Malaysia	19
Philippines	19
Thailand	16
Vietnam	20
Total Banks	**74**

Note: Refer to Appendix 3 for the list of banks segregated by the different countries.

3.6 Definition of Variables

There is currently no consensus derived based on the various study made on this subject on the input and output variable that is to be used to measure the efficiency of banks. The variables that are being used in this study are captured based on Appendix 1, Taxonomy Study on Inputs and Outputs on Banking Article. Taxonomy is not only a tool for systematic storage, efficient and effective teaching/learning and recall for usage of knowledge but it is also a neat way of pointing to knowledge expansion and building (Gattoufi et al., 2004).

The selection of input and output are critical as the efficient DMUs are only efficient in relation to a particular sample and variables combination choice. The efficient units used to measure may not be necessarily deemed efficient in every DEA model combination (Miller & Noulas, 1996).

This study selects three inputs and three outputs. The input variables that have been identified namely, *Deposit & Short Term Funding (DEPO)*, which includes the total customer deposits, deposits from banks and other deposits, and short-term borrowings, *Interest Expense (INTEXP)*, and *Non-Interest Expense (NINTEXP)*. The output variables that have been identified namely, *Total Loan (NLOAN)*, which includes loans to customer and other banks, *Interest Income (INTINC)*, and *Non-Interest Income (NINTINC)*.

The data obtained that is in the home currency of the selected countries as the study using non-parametric do not rely on data belonging to any particular distribution. The non-parametric approach does not require a priori functional specification (Favero & Papi, 1995).

DEA with imprecise data or, more compactly, the Imprecise Data Envelopment Analysis (IDEA) method develop permits mixture of imprecisely and exactly know data which the IDEA models transform into ordinary linear programming forms (Gattoufi et al., 2004).

The bank specific environment variable in this study is identified as the size of banks. Janicky and Prescott (2006) had studied on the size distribution of U.S. banks whereby the findings is that lognormal distribution fits the distribution of bank size. The study had also highlighted that if fits the earlier studies on firm size distribution by Gibrat (1931) which the findings was often known as Gibrat's Law or the Law of Proportionate.

Size is represented by Small (0), Medium (1) and Large (2). The bank size is determined by employing the logarithm of bank total assets as a proxy for absolute bank size and by assumption that the log of bank assets is normally distribution. The grouping of small, medium and large bank are by the percentile grouping whereby small banks is below 33.33 percentile rank, medium banks are between 33.33 to 66.67 and large banks are above 66.67 percentile.

Table 3.2 provides the descriptive statistics that is being used to categorize the bank's size in terms of percentile. The bank size is grouped by country where it is based on the logarithm of total asset of the individual banks in US dollars. Banks with the value of 7.7411 and below are categorized as small banks while banks with value between 7.7412 to 9.4322 are categorized as medium banks and larger than 9.4322 are categorized as large banks.

Table 3.2
Descriptive Statistics to Categorize Bank Size

Statistics		Logarithm of Total Asset mil (USD)
Mean		8.5054
Std. Error of Mean		0.17247
Median		8.5229
Mode		4.18[a]
Std. Deviation		1.48362
Variance		2.201
Skewness		-0.293
Std. Error of Skewness		0.279
Kurtosis		-0.128
Std. Error of Kurtosis		0.552
Range		7.38
Minimum		4.18
Maximum		11.55
Percentile	33.33	7.7411
	66.67	9.4322

The summary of the number of banks that are being categorized as small, medium and large are categorized in Table 3.3 below.

Table 3.3
Bank Size Grouping

Country	Bank Size	Number of Banks
Malaysia	Large	12
	Medium	3
	Small	4
Malaysia Total		**19**
Philippines	Large	3
	Medium	9
	Small	7
Philippines Total		**19**
Thailand	Large	8
	Medium	5
	Small	3
Thailand Total		**16**
Vietnam	Large	3
	Medium	8
	Small	9
	Vietnam Total	**20**
	Total Banks	**74**

Note: Refer to Appendix 4 for the list of banks under the different bank size grouping.

3.7 Data Description

Table 3.4 below presents the descriptive statistic on the data set that is being used for this study. The data that is obtained for the different variables are in the bank's country currency. Based on the data description, it is notable that the range between the minimum amount and the maximum amount is large which is due to the data incorporates small, medium and large banks in the study.

Table 3.4
Descriptive Statistics Segregated by Country

Malaysia - mil (RM)

	Deposit	Interest Expense	Non-Interest Expense	Net Loans	Interest Income	Non-Interest Income
Mean	50195.98	1338.49	845.61	34230.44	2704.71	607.83
Median	34299.00	900.50	585.10	21989.30	1766.00	362.50
Standard Deviation	54567.74	1470.92	1000.81	39371.22	2981.29	737.35
Kurtosis	2.51	3.37	6.66	3.34	2.99	5.83
Skewness	1.72	1.84	2.40	1.87	1.79	2.36
Range	242466.48	6707.10	5548.60	185646.80	13160.40	3644.50
Minimum	665.52	5.30	10.60	136.40	25.20	8.20
Maximum	243132.00	6712.40	5559.20	185783.20	13185.60	3652.70
Sum	4768618.02	127156.70	80332.90	3251891.56	256947.80	57743.80
Count	95	95	95	95	95	95

Philippines - bil (PHP)

	Deposit	Interest Expense	Non-Interest Expense	Net Loans	Interest Income	Non-Interest Income
Mean	169.12	5.32	6.70	82.40	11.92	3.72
Median	101.27	3.88	4.00	42.33	8.36	1.92
Standard Deviation	188.62	5.18	7.39	100.49	12.33	4.51
Kurtosis	1.36	1.43	1.70	2.53	1.01	1.71
Skewness	1.54	1.51	1.56	1.82	1.44	1.61
Range	714.98	19.82	32.03	453.63	48.77	18.84
Minimum	0.09	0.01	0.10	0.09	0.11	-1.20
Maximum	715.08	19.82	32.13	453.72	48.88	17.65
Sum	16066.44	505.70	636.04	7828.09	1132.54	353.07
Count	95	95	95	95	95	95

Thailand - bil (THB)

	Deposit	Interest Expense	Non-Interest Expense	Net Loans	Interest Income	Non-Interest Income
Mean	437.18	9.44	11.71	356.92	25.27	5.68
Median	303.24	6.49	5.55	189.91	16.22	2.22
Standard Deviation	424.49	8.42	11.50	344.22	23.27	6.77
Kurtosis	-0.34	-0.10	-0.86	-0.89	-0.56	0.37
Skewness	0.90	0.86	0.81	0.72	0.82	1.23
Range	1513.59	33.24	34.95	1124.14	80.61	27.63
Minimum	0.12	0.00	0.05	0.14	0.02	-4.72
Maximum	1513.71	33.24	35.00	1124.27	80.62	22.91
Sum	34974.71	755.13	936.75	28553.99	2021.70	454.27
Count	80	80	80	80	80	80

Vietnam - bil (VND)

	Deposit	Interest Expense	Non-Interest Expense	Net Loans	Interest Income	Non-Interest Income
Mean	48313.08	2989.33	938.61	33731.97	4578.82	568.87
Median	18981.90	1067.74	278.90	10871.95	1538.12	148.15
Standard Deviation	75504.20	5200.31	1784.55	60346.17	7842.27	954.88
Kurtosis	8.73	16.61	12.17	12.79	12.72	9.78
Skewness	2.79	3.73	3.37	3.37	3.32	2.89
Range	420744.71	31744.59	9794.20	361386.26	44983.95	5264.71
Minimum	485.54	12.39	7.80	353.49	35.05	1.89
Maximum	421230.25	31756.98	9802.00	361739.75	45019.00	5266.60
Sum	4831307.53	298932.86	93860.50	3373196.74	457882.04	56886.81
Count	100	100	100	100	100	100

Table 3.5 represents the average total deposit, interest expense, non-interest expense, net loans, interest income and non-interest income which have been segregated by country from 2005 to 2009. Based on the data, all the variables selected generally have an increasing trend from 2005 to 2008 however in 2009, the interest expense and interest income have decreased as compared to year 2008 except for Philippines whereby there is a decrease in interest expense but an increase in interest income. This indicates that the banks in the country are able to generate more income with less expense which translates into efficiency in managing its business.

Table 3.5
Data Set Statistics Segregated by Country

Country: Malaysia - mil (RM)						
Year	Deposit	Interest Expense	Non-Interest Expense	Net Loan	Interest Income	Non-Interest Income
2005	38,043.30	940.42	644.66	26,977.30	1,969.72	507.39
2006	45,274.30	1,259.66	733.12	30,668.81	2,494.43	530.31
2007	51,611.52	1,555.11	854.78	33,205.66	2,954.84	643.47
2008	55,083.72	1,628.25	918.72	38,133.34	3,157.09	671.30
2009	60,967.06	1,309.02	1,076.77	42,167.07	2,947.48	686.68

Country: Philippines - bil (PHP)						
Year	Deposit	Interest Expense	Non-Interest Expense	Net Loan	Interest Income	Non-Interest Income
2005	123.23	4.73	5.05	57.04	9.89	2.70
2006	155.54	5.84	6.41	70.24	11.85	4.11
2007	164.45	5.16	6.85	79.42	11.73	4.28
2008	191.79	5.54	7.17	99.70	12.53	3.12
2009	210.59	5.33	8.00	105.60	13.60	4.37

Country: Thailand - bil (THB)						
Year	Deposit	Interest Expense	Non-Interest Expense	Net Loan	Interest Income	Non-Interest Income
2005	393.35	5.70	9.06	316.53	18.71	4.66
2006	419.25	11.81	11.28	334.13	26.78	5.62
2007	429.95	11.89	12.84	352.16	27.96	5.35
2008	464.60	10.49	12.37	389.74	28.30	5.71
2009	478.78	7.32	13.00	392.07	24.61	7.05

Country: Vietnam - bil (VND)						
Year	Deposit	Interest Expense	Non-Interest Expense	Net Loan	Interest Income	Non-Interest Income
2005	26,513.53	1,186.30	473.96	18,953.90	2,111.78	252.22
2006	31,395.03	1,707.22	565.85	22,012.93	2,789.97	322.64
2007	48,499.69	2,544.86	885.57	33,983.82	4,049.26	687.71
2008	58,894.72	5,014.01	1,299.47	39,805.54	7,187.00	758.59
2009	76,262.41	4,494.25	1,468.17	53,903.65	6,756.09	823.18

The Table 3.5 depicts the average amount based on the different variables used segregated by the different country from the period of 2005 to 2009. Based on the above figures, we can summarize that the value for all the different variables are increasing year on year basis except for interest expense and interest income whereby in 2009, the value had decreased as compared to the previous year.

3.8 Normality Testing of Data Set

The data set is being tested for normality test which is essential prior to performing Multiple Regression. A normal distribution is referred as Gaussian distribution and is defined by two parameters that represent the location and scale. The standard normal distribution is deemed normal distributed with a mean of zero and a variance of one.

The normality test is conducted using the Shapiro-Wilk test whereby the dependent and independent variables is being test. The data set are being tested for each country prior to performing Multiple Regression. If the result of 'W' value is close to 1.00, this indicates normality in the data set.

The Shapiro-Wilk test is being performed using Stata (version 10 for Windows) and the results can be referred in Appendix 10.