International Journal on
Advanced Science
Engineering
Information Technology

Vol. 2 (2012) No. 5
ISSN: 2088-5334

Statistical Modelling of CO_2 Emissions in Malaysia and Thailand

Tay Sze Hui[1], Shapiee Abd Rahman[2] and Jane Labadin[3]

Department of Computational Science and Mathematics,
Faculty of Computer Science and Information Technology,
Universiti Malaysia Sarawak, 94300 Kota Samarahan, Sarawak, Malaysia.
E-mail: [1]shtay1011@gmail.com, [2]sar@fit.unimas.my, [3]ljane@fit.unimas.my

***Abstract*— Carbon dioxide (CO_2) emissions is an environmental problem which leads to Earth's greenhouse effect. Much concerns with carbon dioxide emissions centered around the growing threat of global warming and climate change. This paper, however, presents a simple model development using multiple regression with interactions for estimating carbon dioxide emissions in Malaysia and Thailand. Five indicators over the period 1971-2006, namely energy use, GDP per capita, population density, combustible renewables and waste, and CO_2 intensity are used in the analysis. Progressive model selections using forward selection, backward elimination and stepwise regression are used to remove insignificant variables, with possible interactions. Model selection techniques are compared against the performance of eight criteria model selection process. Global test, Coefficient test, Wald test and Goodness-of-fit test are carried out to ensure that the best regression model is selected for further analysis. A numerical illustration is included to enhance the understanding of the whole process in obtaining the final best model.**

***Keywords*— CO_2 emissions; multiple regression; model selection techniques**

I. INTRODUCTION

Carbon dioxide (CO_2) is defined as a colourless, odourless, incombustible and non-poisonous gas produced during combustion of carbon, decomposition of organic compounds and in the respiration of living organisms, as referring to [1]. Carbon dioxide emissions happen when carbon dioxide is released into the atmosphere over a specified area and period of time through either natural processes or human activities. Scientifically, carbon dioxide is a chemical compound that composed of one carbon atom and two oxygen atoms. Much concern with carbon dioxide in particular is that its amount being released has been dramatically increased in the twentieth century. Scientists have found that greenhouse gas emissions such as carbon dioxide possibly contribute to global warming, as pointed out in [2]. CO_2 emissions could aggravate global warming and result in environmental deteriorations and public health problems, as stated in [3]. In the year 2007, the Intergovernmental Panel on Climate Change (IPCC) stated that global average temperatures is likely to increase by between 1.1 and 6.4 ^{o}C during the 21st century [4]. To date, mathematical modelling of carbon dioxide emissions in Malaysia and Thailand is still lacking. Therefore, this study focuses on the modelling of CO_2 emissions in Malaysia and Thailand based on socio-economic and demographic variables using regression analyses.

II. LITERATURE REVIEW

At least until recently, there is clearly a rising awareness about global warming due to man-made mechanical emissions. Thus, there are several efforts being made to analyze CO_2 emissions in different countries or regions of the world. Patterns in CO_2 emissions and its related determinants of many countries or regions of the world have been analyzed in the literature. Reference [5] demonstrated a newly developed dataset involving more than one hundred countries around the world to study the reduced-form relationship between per capita CO_2 emissions and per capita GDP, known as the Environmental Kuznets Curve (EKC). Reference [6] had employed regression models to estimate and compare fuel consumption and CO_2 emissions from passenger cars and buses. Meanwhile, [7] suggested applying decomposition analysis (DA) method on energy-related CO_2 emissions in Greece as well as Arithmetic Mean Divisia Index (AMDI) and Logarithmic Mean Divisia Index (LMDI) techniques on a period-wise and time-series basis. In [8] research, they scrutinized the environmental convergence hypothesis and the stationarity property of relative per capita CO_2 emissions in 21 OECD countries from 1960 to 2000 by using the seemingly unrelated regressions augmented Dickey–Fuller (SURADF) test. Reference [9] examined the relationships between carbon

dioxide emissions, energy consumption and economic growth in China by using multivariate co-integration Granger causality tests. On the other hand, [10] had used a panel vector error correction model to investigate the relationship between carbon dioxide emissions, electricity consumption and economic growth of five ASEAN countries. Reference [3] research had studied on various energy efficiency efforts and carbon trading potential in Malaysia to fight against global warming through reducing greenhouse gases emissions. Based on [11] research, the consumer lifestyle approach of different regions and income levels was used to analyze and explain the impact of carbon dioxide emissions and energy consumption by urban and rural households in China. Reference [12] proposed a dynamic panel data model to examine the determinants of carbon dioxide emissions for a global panel involving 69 countries with the dataset from the year 1985 to 2005. Reference [13] pointed out that applying time series data of a single country only into an investigation may be able to determine and explain past experiences such as energy policies, environmental policies and exogenous shocks.

It is remarkable that most studies are concerned with analyzing the patterns of changes in energy consumption, income and global emissions with those of CO_2 in particular for a range of countries using various methodologies. Despite the increasing sophistication of applications and methodologies employed on a variety of researches, the interrelationship between CO_2 emissions and other variables in Malaysia and Thailand is still lacking and has not been examined extensively up to date. Therefore, this study attempts to provide such an analysis using multiple regression approache. According to [14], multiple regression is the widely used technique when a prediction is needed and where the data on several relevant independent variables are available.

III. Data and Methodology

The data used in this paper are the annual time series data for Malaysia and Thailand from 1971 to 2006. The data were obtained from World Bank's World Development Indicators, as in [15]. The variables employed are CO_2 emissions (metric tons per capita), energy use (kg of oil equivalent per capita), GDP per capita (constant 2000 US$), population density (people per sq. km of land area), combustible renewables and waste (% of total energy), and CO_2 intensity (kg per kg of oil equilavent energy use).

Multiple regression (MR) model is a statistical method used to examine the relationship between a dependent variable and a set of independent variables. Suppose that the value of a dependent variable, Y is influenced by k independent variables, X_1, X_2, X_3, ..., X_k. In general, the multiple regression model is defined as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_k X_k + \varepsilon \qquad (1)$$

where β_0 is the intercept term, β_j denotes the j-th coefficient of independent variable X_j and ε is the random error term. The j-th variables, X_j where j = 1, 2, 3, …, k, can be single independent variables, interaction variables, generated variables, transformed variables or dummy variables. The regression coefficients were estimated using ordinary least square (OLS) method in order to obtain a model that would describe the data, as stated in [16].

There are some basic assumptions of multiple regression which must be statisfied so that the results will not be biased. The assumptions are:

a) The error term, ε has a zero mean value for any set of values of the independent variables such that $E(\varepsilon) = 0$.
b) Homoscedasticity, that is the variance of ε, is constant such that $var(\varepsilon) = \sigma^2$.
c) The error term, ε follows the normal distribution with zero mean and variance σ^2 such that $\varepsilon \sim N(0, \sigma^2)$.
d) The error term, ε is uncorrelated with one another such that their covariance is zero, $cov(\varepsilon_i, \varepsilon_j) = 0$ for $i \neq j$. It means that there is no autocorrelation exists between the error terms.
e) No exact collinearity or no multicollinearity exists between the k independent variables.

The regression model with k variables and k+1 parameters including the constant term as expressed in equation (1) is one of the possible models. All the possible models are listed out based on single independent variables and all possible interactions of related single independent variables either generated or transformed. If multicollinearity phenomenon exists, then the source variables in the regression models are removed. In order to obtain appropriate regression models, Global test and Coefficient test are conducted to test the overall statistical siginificance of the independent variables on the dependent variable, as in [17]. Then the regression models after the final elimination are the selected models free from problems of multicollinearity and insignificance. This process is known as data-based model simplification.

The process of selecting a subset of original predictive variables is by means of removing variables that are either redundant or with little predictive information, as in [18]. Thus, it is useful to enhance the comprehensibility of the resulting models so as to generalize better. There are three popular optimization strategies employed in model selection, namely forward selection, backward elimination and stepwise regression. In this study, the model selection algorithm is performed by using PASW Statistics Software. Forward selection starts with an empty set of variable and gradually adds in variables that most improve the model performance until there is no additional variable that satisfies the predetermined significance level. By contrast, backward elimination method begins with a full set of all individual variables and sequentially eliminates the least important variable from the model. The process ends when an optimum subset of variables is found. As for stepwise regression, it is a combination of forward selection and backward elimination that determines whether to include or exclude the individual variables at each stage. The variable selection terminates when the measure of all variables in the variable set is maximized.

Reference [16] had also explained in detail the statistical procedures of obtaining the best model based on model selection criteria. The model selection criteria are Akaike information criterion (AIC), finite prediction error (FPE), generalised cross validation (GCV), Hannan and Quinn criterion (HQ), RICE, SCHWARZ, sigma square (SGMASQ) and SHIBATA. The whole selection criteria is based on the

residual sum of squares (RSS) multiplied by a penalty factor which would depend on the model complexity. Model with higher complexity generally will decrease the RSS but increase the penalty. These criteria thus allow trade-offs between goodness-of-fit and model complexity. The model with the lowest values for most of the criteria statistics is preferable and chosen as the best model. The joint significances of regression coefficients are examined by the Wald test, followed by the goodness-of-fit test so as to investigate the suitability of the final model.

IV. Results and Discussions

CO_2 emissions (Y) as the dependent variable was related to energy use (X_1), GDP per capita (X_2), population density (X_3), combustible renewables and waste (X_4), and CO_2 intensity (D). In this study, only the data for population density was normally distributed in its level form. Since the data for other variables were not normally distributed, they were transformed into natural logarithms prior to analysis because this helps to induce normality. Meanwhile, CO_2 intensity was generated into dummy variable since it was still not normal after several transformations.

Table I demonstrates the relationship between CO_2 emissions and the determinants that are related. There is a significant relationship between the variable X_1, X_2, X_4 and D. It is obvious that the energy use (X_1), GDP per capita (X_2) and combustible renewables and waste (X_4) are highly correlated with the carbon dioxide emissions (Y). Furthermore, a positive significant relationship exists between Y and X_1 (r = 0.9773, p-value < 0.01), Y and X_2 (r = 0.9806, p-value < 0.01) as well as Y and D (r = 0.6166, p-value < 0.01). From the highlighted triangle shown in Table I, there exists multicollinearity such that the absolute value of the correlation coefficient is greater than 0.95 among the independent variables. Hence, the multicollinearity source variables have to be removed from the model. After resolving the multicollinearity problem, further analysis can then be carried out.

All the possible models are subjected to Global test and Coefficient test. For illustration purpose, model BM31.10, the backward elimination model 31 after 10 times of the multicollinearity source variable removals, was considered. Table II represents the ANOVA table for Global test. The hypothesis of Global test for model BM31.10 is as follows:

H_0: $\beta_4 = \beta_{12} = \beta_{34} = \beta_{123} = \beta_{124} = \beta_{1D} = \beta_{3D} = \beta_{4D} = 0$
H_1: At least one of the β's in H_0 is nonzero.

From Table II, the F_{cal} is 2726.85 and the $F_{critical}$ is $F_{0.05, 8, 63} = 2.10$. Since F_{cal} is greater than $F_{critical}$, the decision is to reject the null hypothesis where all the regression coefficients in model BM31.10 are zero.

TABLE I
A Pearson Correlation Table Between CO2 Emissions and Its Determinants

	Y	X_1	X_2	X_3	X_4	D
Y	1	0.9773(**) 0.0000	0.9806(**) 0.0000	-0.0147 0.9026	-0.9039(**) 0.0000	0.6166(**) 0.0000
X_1	0.9773(**) 0.0000	1	0.9707(**) 0.0000	-0.0059 0.9608	-0.9189(**) 0.0000	0.4973(**) 0.0000
X_2	0.9806(**) 0.0000	0.9707(**) 0.0000	1	-0.1551 0.1934	-0.9542(**) 0.0000	0.5078(**) 0.0000
X_3	-0.0147 0.9026	-0.0059 0.9608	-0.1551 0.1934	1	0.3845(**) 0.0009	0.1873 0.1151
X_4	-0.9039(**) 0.0000	-0.9189(**) 0.0000	-0.9542(**) 0.0000	0.3845(**) 0.0009	1	-0.3875(**) 0.0008
D	0.6166(**) 0.0000	0.4973(**) 0.0000	0.5078(**) 0.0000	0.1873 0.1151	-0.3875(**) 0.0008	1

** Correlation is significant at the 0.01 level (2-tailed).

TABLE II
The ANOVA Table for Global Test

Source of Variations	Sum of Squares	df	Mean Square	F
Regression	7.3431	8	0.9179	2726.85
Residual	0.0212	63	0.0003	
Total	7.3643	71		

The best model for CO_2 emissions estimation is selected by first applying the backward elimination method. Then, the Coefficient test is carried out for all the coefficients in the model where Table III shows the coefficient for each variable of the model BM31.10.3 with the last digit is the number of insignificant variables being eliminated.

The criteria condition used in this regression analysis is by dropping the variable with the p-value > 0.05. From the observations in Table III, the variable X_3, X_{34} and X_1D are removed from the regression model since their p-values are greater than 0.05. It indicates that the corresponding variables are insignificant at $\alpha = 0.05$. The resulting model contains only significant variables with all the p-values less than 0.05. Similar procedures are applied to the forward selection and stepwise regression method for model selection. After progressive eliminations, the final model is thus obtained and expressed as in equation (2).

$$Y = -0.3728 - 0.6769X_4 + 0.0885X_{12} + 0.0001X_{123} + 0.0481X_{124} - 0.0006X_3D + 0.1029X_4D \quad (2)$$

The Wald test is performed on the final model where the unrestricted model denoted as (U) and the restricted model denoted as (R) are expressed respectively in the equation (3) and (4) as follows:

$$(U): Y = \beta_0 + \beta_4X_4 + \beta_{12}X_{12} + \beta_{34}X_{34} + \beta_{123}X_{123} + \beta_{124}X_{124} + \beta_{1D}X_1D + \beta_{3D}X_3D + \beta_{4D}X_4D + \varepsilon \quad (3)$$

$$(R): Y = \beta_0 + \beta_4X_4 + \beta_{12}X_{12} + \beta_{123}X_{123} + \beta_{124}X_{124} + \beta_{3D}X_3D + \beta_{4D}X_4D + \varepsilon \quad (4)$$

The hypothesis of Wald test is:

H_0: $\beta_{34} = \beta_{1D} = 0$
H_1: At least one of the β's in H_0 is nonzero.

As shown in Table IV, F_{cal} is 1.5753 and $F_{critical}$ is $F_{0.05, 2, 63} = 3.15$. The decision is not to reject the null hypothesis where all the eliminated regression coefficients are zero since F_{cal} is less than $F_{critical}$. Thus, this justifies the removal of the insignificant variables in the coefficient test. In order to select the best model from forward, backward and stepwise selection method, the model selection criteria process is conducted. The models to be compared with are shown in Table V, namely forward selection model (FM26.8.3), backward elimination model (BM31.10.3) and stepwise regression model (SM31.10.3). Majority of the criteria indicates that BM31.10.3 and SM31.10.3 are the two best models for CO_2 emissions as both models show similar findings with the same regression equation as expressed in (2).

TABLE III
THE COEFFICIENTS IN MODEL BM31.10.3

Model BM31.10.3	Unstandardized Coefficients		t-values	p-values
	B	Std. Error		
Constant	-0.3728	0.2602	-1.4329	0.1567
X_4	-0.6769	0.0993	-6.8187	0.0000
X_{12}	0.0885	0.0198	4.4709	0.0000
X_{123}	0.0001	0.0000	4.2676	0.0001
X_{124}	0.0481	0.0049	9.8776	0.0000
X_3D	-0.0006	0.0002	-2.7122	0.0085
X_4D	0.1029	0.0130	7.8870	0.0000

Excluded Variables(b)

Model BM31.10.3	Beta In	t-values	p-values	Partial Correlation	Collinearity Statistics Tolerance
X_3	-0.0071(a)	-0.0624	0.9504	-0.0078	0.0036
X_{34}	-0.0129(a)	-0.1674	0.8676	-0.0209	0.0080
X_1D	0.0635(a)	1.6631	0.1012	0.2035	0.0310

a. Predictors in the Model: Constant, X_4D, X_{12}, X_{124}, X_3D, X_{123}, X_4
b. Dependent Variable: Y

TABLE IV
THE WALD TEST

Source of Variations	Sum of Squares	df	Mean Square	F
Differences	0.0011	2	0.0005	1.5753
Unrestricted (U)	0.0212	63	0.0003	
Restricted (R)	0.0223	65		

TABLE V
THE MODEL SELECTION CRITERIA FOR THE CORRESPONDING MODELS

Model	RSS	AIC	FPE	GSC	HQ	RICE	SCHWARZ	SGMASQ	SHIBATA
FM26.8.3	0.0402	0.0006	0.0006	0.0005	0.0007	0.0006	0.0008	0.0006	0.0006
BM31.10.3	0.0223	0.0004	0.0004	0.0003	0.0004	0.0004	0.0005	0.0003	0.0004
SM31.10.3	0.0223	0.0004	0.0004	0.0003	0.0004	0.0004	0.0005	0.0003	0.0004

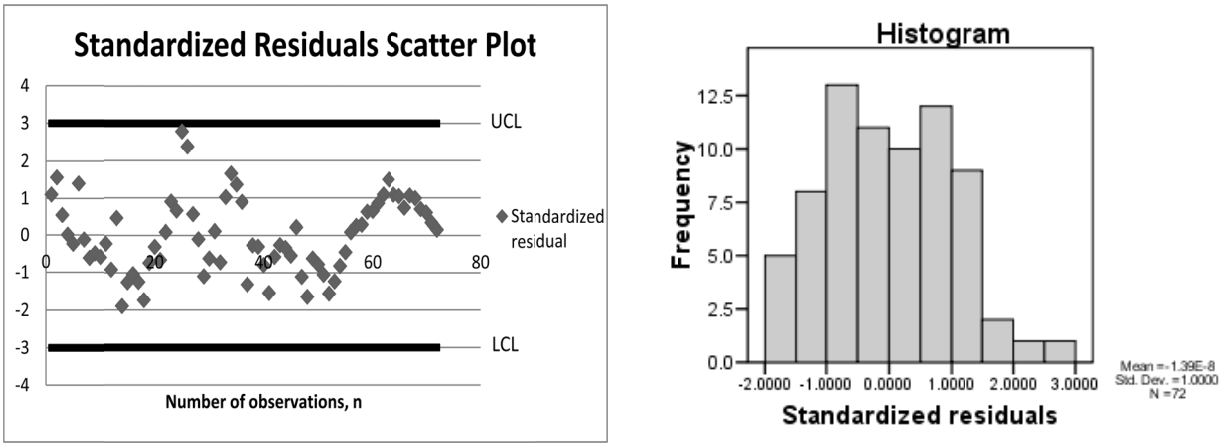

Fig. 1 The scatter plot and histogram for standardized residuals

Based on the best model, the standardized residuals are obtained and goodness-of-fit test is carried out. The scatter plot in Fig. 1 shows that the standardized residuals are randomly distributed since there is no obvious pattern observed. In addition, the normality test using Kolmogorov-Smirnov statistics has shown that the standardized residuals are distributed normally with zero mean, standard deviation approximates to 1 and the *p*-value is 0.200. Thus, it means that the best model is a well represented model in describing the carbon dioxide emissions.

By substituting all the data entry needed for the estimated model in the equation (2), the estimated CO_2 emissions for the year 2007 is obtained. In 2007, the actual value after transformation for CO_2 emissions in Malaysia and Thailand are 0.8643 and 0.6170 respectively. The actual value is compared with the estimated CO_2 emissions, that is, 0.8292 for Malaysia and 0.6159 for Thailand. It is found that the difference between actual and estimated CO_2 emissions is 4.06% for Malaysia and 0.17% for Thailand. Since the variation is quite small, it can be concluded that the estimated model is suitable to predict the future carbon dioxide emissions.

V. CONCLUSION

The best model in this study is found to be either the multiple regression model BM31.10.3 using backward elimination or SM31.10.3 using stepwise multiple regression. The combustible renewables and waste (X_4), which comprise solid biomass, liquid biomass, biogas, industrial waste and municipal waste, is the only main determinant that influences the CO_2 emissions in both Malaysia and Thailand. The negative regression coefficient shows that the CO_2 emissions will be reduced whenever there is an increase in the variable X_4. This implies that when the countries use more combustible renewables and waste to generate energy and electricity, the CO_2 emissions will be decreased and it leads to less pollution. Other independent variables cannot act as a single-effect variable since these variables do not have a direct effect on the CO_2 emissions. However, they interact together to indicate the strength of contribution in determining the occurrence of CO_2 emissions, for instance, the variable X_1X_2 indicates that the energy use interacts with the GDP per capita. Since there exists effect of higher order interactions, the polynomial regression could be considered in future studies. Besides that, other relevant determinants such as trade openness, per capita income, energy intensity and electricity consumption could also be included in the model.

REFERENCES

[1] OECD. 2011. Glossary of Statistical Terms. [Online]. Available: http://stats.oecd.org/glossary/ [21 May 2011]

[2] M. Lanne and M. Liski, "Trends and Breaks in Per-capita Carbon Dioxide Emissions," *Energy Journal*, vol. 25, pp. 41–65, 2004.

[3] T. H. Oh and S. C. Chua, "Energy Efficiency and Carbon Trading Potential in Malaysia," *Renewable and Sustainable Energy Reviews*, vol. 14, pp. 2095–2103, 2010.

[4] Intergovernmental Panel on Climate Change (IPCC), *Climate Change 2007: Synthesis Report*. Geneva, Switzerland: IPCC, 2007.

[5] M. Galeottia and A. Lanza, "Desperately Seeking Environmental Kuznets," *Environmental Modelling & Software*, vol. 20, pp. 1379–1388, 2005.

[6] J. A. Paravantis and D. A. Georgakellos, "Trends in Energy Consumption and Carbon Dioxide Emissions of Passenger Cars and Buses," *Technology Forecast Social Change*, vol. 74, pp. 682–707, 2007.

[7] E. Hatzigeorgiou, H. Polatidis, and D. Haralambopoulos, "CO_2 Emissions in Greece for 1990–2002: A Decomposition Analysis and Comparison of Results Using the Arithmetic Mean Divisia Index and Logarithmic Mean Divisia Index Techniques," *Energy*, vol. 33, pp. 492–499, 2008.

[8] C. C. Lee and C. P. Chang, "New Evidence on the Convergence of Per Capita Carbon Dioxide Emissions from Panel Seemingly

Unrelated Regressions Augmented Dickey– Fuller Tests," *Energy*, vol. 33, pp. 1468–1475, 2008.

[9] C. C. Chang, "A Multivariate Causality Test of Carbon Dioxide Emissions, Energy Consumption and Economic Growth in China," *Applied Energy*, vol. 87, pp. 3533–3537, 2010.

[10] H. L. Hooi and S. Russell, "CO_2 Emissions, Electricity Consumption and Output in ASEAN," *Applied Energy*, vol.87,pp 1858–1864, 2010.

[11] Z. H. Feng, L. L. Zou, and Y. M. Wei, "The Impact of Household Consumption on Energy Use and CO_2 Emissions in China," *Energy*, vol. 36, pp. 656–670, 2011.

[12] S. S. Susan, "Determinants of Carbon Dioxide Emissions: Empirical Evidence from 69 Countries,"*Applied Energy*, vol.88, pp.376–382, 2011.

[13] D. I. Stern, M. S. Common, and E. B. Barbier, "Economic Growth and Environmental Degradation: The Environmental Kuznets Curve and Sustainable Development," *World Development*, vol. 24, pp. 1151–1160, 1996.

[14] K. Nikolopoulos, P. Goodwin, A. Patelis, and V. Assimakopoulos, "Forecasting with Cue Information: A Comparison of Multiple Regression with Alternative Foreacasting Approaches," *European Journal of Operational Research*, vol. 180, pp. 354–368, 2007.

[15] World Bank. 2011. World Development Indicators. [Online]. Available: http://data.worldbank.org/ [15 March 2011]

[16] R. Ramanathan, *Introductory Econometrics with Applications*, 5th ed. Ohio, United States: Thomson Learning Ohio, 2002.

[17] D. A. Lind, W.G. Marchal and R. D. Mason, *Statistical Techniques in Business & Economics*, 11th ed. New York, United States: McGraw Inc., 2005.

[18] Y. S. Kim, "Towards a Successful CRM: Variable Selection, Sampling and Ensemble," *Decision Support Systems*, vol. 41, pp. 542–553, 2006.