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 ^oC 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