Comparison between Multiple Regression and Multivariate Adaptive Regression Splines for Predicting CO_2 Emissions in ASEAN Countries

Tay Sze Hui, Shapiee Abd Rahman and Jane Labadin
Department of Computational Science and Mathematics,
Faculty of Computer Science and Information Technology,
Universiti Malaysia Sarawak, 94300 Kota Samarahan, Sarawak, Malaysia.
shtay1011@gmail.com, {sar, ljane}@fit.unimas.my

***Abstract*—Global warming due to the rapid increase in greenhouse gas emissions, mainly carbon dioxide (CO_2), is a worldwide issue that leads to escalating pollutions and emerging diseases. The comparative performances of multiple regression (MR) and multivariate adaptive regression splines (MARS) for statistical modelling of CO_2 emissions are analyzed in ASEAN countries over the period of 1980-2007. The regression models are fitted individually for every potential variable investigated so as to find the best-fit parametric or non-parametric model. The results show a significant difference between the performance of MR and MARS models with the inclusion of interaction terms. The MARS model is computationally feasible and has better predictive ability than the MR model in predicting CO_2 emissions. In overall, MARS can be viewed as a modification of stepwise regression that enhances the latter's performance in the regression setting.**

Keywords—CO_2 emissions; ASEAN; multiple regression; multivariate adaptive regression splines

I. INTRODUCTION

Carbon dioxide (CO_2) is an abundant greenhouse gas in the atmosphere besides methane, nitrous oxide, ozone and water vapour. The dramatic increase in CO_2 atmospheric concentrations is primarily due to the combustions of fossil fuels and the changes in land use practices since the Industrial Revolution. As a result, global average surface warming and sea level would continue to rise over the time scales associated with climatic processes and impacts [1].

The enhanced greenhouse effect is attributable to the higher levels of trapped thermal infra-red radiation which have been directed outward from the Earth [2]. Different greenhouse gases, which are responsible for global warming, have their own heat-trapping abilities. The gas responsible for the most warming is carbon dioxide. Other greenhouse gases do not cause as much warmth to the atmosphere as carbon dioxide does as the concentrations of these gases are much lower than carbon dioxide [3]. The world's carbon footprint has increased by over a third since 1998 [4].

The primary aim of this study is to develop the regression models using multiple regression and multiple adaptive regression splines techniques. The secondary aim is to compare the model performance for predicting CO_2 emissions. This is to determine whether a parametric or non-parametric regression model is more appropriate in describing CO_2 emissions for ASEAN countries.

II. DATA

In this study, the data employed are the annual data of carbon dioxide emissions (Y), energy consumption (X_1), gross domestic product (X_2), carbon intensity (X_3), energy intensity (X_4), population (X_5), crop production index (X_6) and income group (D). The dataset cover all the ASEAN countries from year 1980-2007 which are obtained from World Bank's World Development Indicators, Energy Information Administration and World Economic Outlook. The data are normalized into the interval [0, 1] to avoid instabilities that could affect the quality of the final model due to the widely different locations and scales of variables [5]. The dataset is divided into two sub-samples, namely a training set (1980-2000) and a testing set (2001-2007). The training set is used to fit the model while the testing data is used to validate the model.

III. METHODOLOGY

The modelling studies have developed different statistical data mining techniques to analyze the relationships between dependent variable and independent variables and unravel the complexity of interactions among the independent variables. Those data mining methodologies included with regression algorithms such as multiple regression and multivariate adaptive regression splines are implemented in this study.

A. Multiple Regression (MR)

The term "multiple regression" was introduced by [6]. The general multiple regression model is defined in (1) as

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

The dependent variable, Y, is related to k independent variables, X_1, X_2, ..., X_k. In this statistical model, the intercept term is denoted as β_0, the coefficients of independent variables are represented by β_1, β_2, ..., β_k and ε refers to the error term. In this study, all subsets model selection up to the first order of interactions using stepwise regression search algorithm is used to determine the significant independent variables. Generally, the basic assumptions of multiple regression models are:

978-1-4799-1092-2/13/$31.00 ©2013 IEEE

a) The error term has a zero mean value for any set of values of the independent variables such that $E(\varepsilon) = 0$.

b) The error term achieves homoscedasticity, that is the variance of ε is constant, such that $var(\varepsilon) = \sigma^2$.

c) The error term follows the normal distribution with zero mean and variance σ^2 such that $\varepsilon \sim N(0, \sigma^2)$.

d) The error terms are independent of one another such that ε_i and ε_j are independently distributed for all $i \neq j$.

e) Each of the independent variable, X_j is uncorrelated with the error term such that their covariance, $cov(X_j, \varepsilon_j) = 0$.

B. Multivariate Adaptive Regression Splines (MARS)

Multivariate adaptive regression splines (MARS) is a powerful mathematical regression tool proposed by Friedman [5] with the aim to predict the value of a dependent variable from a set of independent variables. In particular, MARS approach is very suitable to be used in data mining, especially when handling fairly large datasets since the spline functions will automatically divide the input data into various sub-regions [7]. MARS does not require any a priori assumptions about the underlying functional relationship between the dependent and independent variables. The basis functions of MARS are splines which are commonly used to build flexible regression models because of their compact support and smoothness properties [8]. A truncated spline function consists of a left-sided and a right-sided segment defined by a given knot location *t*, as shown in (2) and (3) respectively.

$$b_q^-(x-t) = [-(x-t)]_+^q = \begin{cases} (t-x)^q & if\ x < t \\ 0 & otherwise \end{cases} \quad (2)$$

$$b_q^+(x-t) = [+(x-t)]_+^q = \begin{cases} (x-t)^q & if\ x \geq t \\ 0 & otherwise \end{cases} \quad (3)$$

It is to be noted that *q* is the power to which the splines are raised in order to determine the degree of smoothness of the resultant function estimate. The subscript "+" in the function indicates that the output is zero when the argument is unsatisfied. The general MARS model is represented in (4):

$$\hat{y} = a_0 + \sum_{m=1}^{M} a_m B_m(x) \quad (4)$$

where $\hat{y}$ is the predicted value for the dependent variable, a_0 is the coefficient of the constant basis function, *M* is the number of basis functions included in the MARS model, a_m is the coefficient of the *m*th basis function and $B_m(x)$ is the *m*th basis function which can be either a single spline function or an interaction between two or more spline functions.

The generalized cross-validation (GCV) criterion is used to find the overall best model in MARS analysis [9], as defined in (5):

$$GCV(M) = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y})^2}{\left[1 - \frac{C(M)}{n}\right]^2} \quad (5)$$

where *n* is the number of observations in the data set, y_i is the dependent variable for observation *i*, $\hat{y}$ is the predicted dependent variable and *C(M)* is a complexity penalty function used to avoid over-fitting and increase the model parsimony. *C(M)* is usually defined as in (6):

$$C(M) = M + cd \quad (6)$$

where *M* is the number of non-constant terms in the model, *c* is an user-defined cost penalty factor for each basis function optimization and *d* is the effective degrees of freedom which equals to the number of independent basis functions in the model.

In general, there are three main steps involved in MARS analysis. The first step is a constructive phase, in which the variable with the selected pair of spline functions is introduced and new spline functions are added stepwise. In this procedure, a complex multivariate model named the global MARS model is built where over-fitting problem usually occurs. The second step is the pruning phase, in which some spline functions of the global MARS model are eliminated stepwise using a sequence of general cross-validations alternated with ten-fold cross validation. In the third step of MARS analysis, the optimal model is selected from a series of less complex models using a cross-validation technique.

C. Model Evaluation and Validation

Model evaluation is important in modelling process as to examine whether a statistical model built can describe the system accurately. Eventually, the final model of MARS is compared against the best-fit MR model obtained. Wilcoxon signed-rank test is used to test and compare the results of the prediction modelling techniques in terms of absolute residual error (ARE) and magnitude of relative error (MRE) [10]. It is a distribution-free method that does not rely on the data relating to any underlying distributions. Besides, the test statistic is based on the signs and ranks of observations, not the magnitude; hence it is not affected by the outliers.

The initial step for Wilcoxon signed-rank test is computing the differences between two related observations. The observations with no difference are omitted from the sample. Then, the remaining absolute differences are ranked from smallest to largest, employing tied observations where appropriate. A mean rank is assigned to each tied observation. The signs of each difference are affixed to their corresponding ranks. Subsequently, the sums of both positive and negative ranks are calculated and compared. The sum of the ranks having a plus sign is denoted as T_+ while the sum of the ranks having a minus sign is denoted as T_-. The smaller of these two rank sums is used as the test statistic and referred to as *T*.

When the sample size is large, the distribution of *T* is closely approximated by a normal distribution with a mean of