Fuel 177 (2016) 274–278

Contents lists available at ScienceDirect

Fuel

journal homepage: www.elsevier.com/locate/fuel

Estimation of coal gross calorific value based on various analyses by random forest method

S.S. Matin [a], S. Chehreh Chelgani [b,*]

[a] Department of Environment and Energy, Science and Research Branch, Islamic Azad University, Tehran, Iran
[b] Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48109, USA

HIGHLIGHTS

- Properties of US coal were studied for the prediction of gross calorific value (GCV).
- Random forest (RF) models indicated that RF can accurately predict GCV.
- RF models are much suitable to assess complicated relationships in coal processing.
- Results recommended random forest as a model can be applied for other coal resources.

ARTICLE INFO

Article history:
Received 12 February 2016
Received in revised form 4 March 2016
Accepted 10 March 2016
Available online 16 March 2016

Keywords:
Gross calorific value
Random forest
Proximate analysis
Ultimate analysis
Regression

ABSTRACT

The last decade has witnessed of increasing the application of random forest (RF) models that are known as an exhibit good practical performance, especially in high-dimensional settings. However, on the theoretical side, their predictive ability markedly remains unexplained, especially in coal preparation. RF as a predictive model can tend to work well with large dimensional databases and rank predictors through its inbuilt variable importance measures. In this study, relationships among ultimate and proximate analyses of 6339 US coal samples from 26 states with gross calorific value (GCV) have been investigated by multivariable regression (MVR) and random forest (RF) models. RF method has been used for the variable importance. Models have shown that the ultimate analysis parameters are the most suitable estimators for GCV and that RF can predict GCV quite satisfactory. Running of the best arranged RF structures for the input sets and assessment of errors have suggested that RF models are suitable for complicated relationships.

1. Introduction

Consumption of fossil fuels continues to grow, although there are signs that the rate of growth may be slowing [1]. Also the abundance and versatility of coals makes a vital role for it in nowadays industrial fields like cement making and conversion to coke for the smelting of iron ore [2,3].

The calorific value (CV) as an indicator of the chemically stored energy in coal is a very important parameter in the assessment of its value as a fuel [4,5], and potentially could be a basis for the purchase of coal [6]. CV (heating value) is the amount of energy per unit mass released upon complete combustion [7]. CV of coal as a rank parameter depends on the maceral and mineral composition has a great importance in the conversion of coal to other useful forms of fuel [6]. In detail, CV is defined as the amount of heat evolved when a unit weight of the fuel is burnt completely in oxygen (in a calorimetric bomb) under the conditions specified and the combustion products cooled to a standard temperature of 298 K [3] and called it "gross calorific value (GCV)" that is typically expressed as Btu (or MJ/kg) [8,9].

There are many factors that potentially could cause of errors in GCV determination; obtaining a representative sample of such a heterogeneous material as coal, lack of reproducibility among samples and among analysts, incomplete combustion of the sample, varied the combustion temperature based on coal rank, and the mineral matter content and the nature of their reactions [5]. Thus, estimation of GCV based on the composition of fuel could be one of the basic steps in performance modeling and calculations on combustion systems and evaluate errors of analysis [10].

The main advantage of the GCV prediction is; providing an easy and quick estimation that saving the efforts involved in the experimental measurements [3,11]. In all thermal power plants, the determination of calorific value, proximate analysis, and ultimate analysis are common practice to assess the quality of coals [12]. Hence, there are a number of equations (multivariable regressions

* Corresponding author.
E-mail address: schehreh@umich.edu (S.C. Chelgani).

http://dx.doi.org/10.1016/j.fuel.2016.03.031
0016-2361/

(MVR)) have been developed for the GCV prediction based on proximate and ultimate analyses of various coal samples from different deposits [3,4,10–22].

These models could be used through the investigation of the coal-quality influence on the fuel during the process performance. Equations are mainly linear (although the relationship among the GCV and a few constituents could be nonlinear) [3]. A disadvantage of applying MVR is that it could be subjected to a priori assumption of function form before regression. The function forms may lead to inaccurate or even absurd results if assumed improperly [23,24]. Moreover, MVR models are not suitable for applications that the inter-correlation among predictors and the response has a complex nature, including higher-order interactions and correlations among predictors [25]. Many soft computing methods, such as black box base models, (artificial neural network (ANN), Nero-fuzzy (such as Adaptive neuro-fuzzy inference system (ANFIS)), and genetic algorithm (GA)) have been developed to overcome this problem for the GCV prediction [3,4,8,17,26,27]. These methods are capable of capturing complex relationships among large numbers of variables and have chiefly been used to predict the values of target variables, but they do not necessarily give any particular or explicit insight into the relationships between inputs and the target variables. This is a major drawback in situations where such information would be very important. This has led to the development of so-called variable importance algorithms that can be used in conjunction with black box models to identify the individual effects of explanatory variables [28]. In modeling for prediction of a parameter, measuring the importance of predictors, or estimating of the correlation among variables can be an essential key to have a successful prediction [25,29].

Random forest (RF) models are a recent method which can overcome the latter problems by providing attractive addition to nonlinear approximation of statistical relationships among variables [30]. RF is a model ensemble method built based on combines several decisions by the regression and classification trees (CART, [31]). RF excels in predictive modeling, providing a unified way of defining distance for data with a mixture of continuous and categorical variables [25,32]. Where RF variables are both continuous and categorical, it has several advantages over other statistical modeling methods; based on a large number of trees RF produces low-bias and low-variation results with highly accurate classification and good predictions [29]. However, the main advantage of RF methods is that they tend to work well with larger and higher dimensional data which is able to rank predictors through its inbuilt variable importance measures [25,32]. There is a growing literature on using tree-based methods to model complex relationships such as RF [33–36]. The results have been obtained showed the relative superiority of the RF method over MVR and ANN [37]. But variable importance analyses associated with nonlinear models in general, and RFs in particular, have yet to appear meaningfully in the coal processing literature. The aim of the present work is the assessment of properties of 6339 US coal samples from 26 different states to estimate the GCV as a coal rank parameter and possible variations with respect to ultimate and proximate analysis by using multivariable regression, SPSS software package, and the RF, "R" software package. To our best knowledge, no tree or RF based methods have been proposed for the estimation of coal rank parameters.

2. Materials and methods

2.1. Database

A remarkable computing model requires a comprehensive database to cover a wide variety of conditions. The model should be capable of estimating variables with a high validity. In this study data used to examine the proposed approaches provided from U. S. Geological Survey Coal Quality (COALQUAL) database, open file report 97-134 [38]. The samples with more than 25% ash as well as the samples with proximate and/or ultimate analysis different from 100 were excluded from the database (no US coals with greater than 25% ash are used in power production). A total of 6339 set of coal sample analysis were used. The database, including the determined proximate and ultimate analysis as well as calorific value in as received basis, is also given in the supplementary database to this manuscript.

The procedures of sampling and analytical chemical methods can be found on the http://energy.er.usgs.gov/products/databases/CoalQual/index.htm web address. The number of samples and range of GCV for different states are shown in Table 1. According to the ASTM standard proximate analysis, fixed carbon (FC) is calculated based on other variables (FC% = 100 − (moisture + volatile matter + ash)) and; therefore, FC gathers all errors of other lab tests (moisture, volatile matter, and ash). Therefore, it is not necessary to use all four parameters since, by definition, the four parameters are a closed system, adding to 100% [39]. For the same reason oxygen (oxygen = 100 − (hydrogen + nitrogen + sulfur + carbon)) from ultimate analysis did not consider as a predictor for the GCV modeling. From the total database used in the modeling, randomly 70% of samples are selected for training phase and 30% of data for testing phase of models. Selected samples for the both models (MVR and RF) are exactly the same for a potential comparison.

2.2. Multivariable regression

To choose the most effective predictors for a model in regression, inter-correlation could be used to explore relationships among all inputs and outputs. Inter-correlation is a term applied to denote the correlation of a number of variables among themselves, as distinct from the correlations between them and outputs. The inter-correlation between two variables reflects the degree to how the variables are related. The most common measure of correlation is the Pearson Product Moment Correlation (called Pearson's correlation for short "*r*"). Pearson correlation (inter-correlation) is a measure of linear association between two variables. *r* values range from +1 to −1. The sign of the correlation indicates the direction of the relationship, and its absolute value indicates the strength, with larger absolute values indicating stronger relationships. A negative value for the correlation implies a negative or inverse association, where a positive value means a positive association [40]. Table 2 indicates the Pearson correlation through the entire database.

Stepwise multivariable regression was used to model the relationship among input variables to predict GCV. In the stepwise modeling, variables are sequentially entered into the equation. The first variable considered for reflecting into the equation is the one with largest positive or negative correlation with the dependent variable. This variable is entered into the equation only if it satisfies the criterion for entry. The next variable, with the largest partial correlation, is considered as the second equation input. The procedure stops when there are no variables that meet the entry criterion [40].

2.3. Random forest

As can anticipate the name "Random Forest", a RF model is an ensemble of many classification or regression trees designed to produce accurate predictions which do not over-fit the data [30]. In RF method, variable importance analyses associated with nonlinear. An approach to quantifying the importance of variables in function approximation is to permute the values of the predictor variables, one at a time, and determine the decrease in model accu-