Measurement 94 (2016) 416–422

Contents lists available at ScienceDirect

Measurement

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

# Modeling of free swelling index based on variable importance measurements of parent coal properties by random forest method

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

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

ARTICLE INFO

*Article history:*
Received 27 May 2016
Received in revised form 8 July 2016
Accepted 25 July 2016
Available online 26 July 2016

*Keywords:*
FSI
Random forest
Variable importance
Conventional analyses

ABSTRACT

Coke quality has a critical role in the steelmaking industry. The aim of this study is to examine the complex relationships between various conventional coal analyses using coke making index "free swelling index (FSI)". Random forest (RF) associated with variable importance measurements (VIMs), which is a new powerful statistical data mining approach, is utilized in this study to analyze a high-dimensional database (3961 samples) to rank variables, and to develop an accurate FSI predictive model based on the most important variables. VIMs was performed on various types of analyses which indicated that volatile matter, carbon, moisture (coal rank parameters) and organic sulfur are the most effective coal properties for the prediction of FSI. These variables have been used as an input set of RF model for the FSI modeling and prediction. Results of FSI model indicated that RF can provide a satisfactory prediction of FSI with the correlation of determination $R^2$ = 0.96 and mean square error of 0.16 from laboratory FSIs (which is smaller than the interval unit of FSI; 0.5). Based on this result, RF can be used to rank and select effective variables by evaluating nonlinear relationships among parameters. Moreover, it can be further employed as a non-parametric reliable predictive method for modeling, controlling, and optimizing complex variables; which to our knowledge has never been utilized in the fuel and energy sectors.



## 1. Introduction

Coke has been the source of energy for around 70% of total steel production in the world. In 2014, the European Commission published an updated list of 14 new critical materials (metals, groups of metals or minerals) that are essential for various regional industries, of which coking coal is a critical component [1]. The steelmaking industry controls an essential portion of energy consumption in many countries as steel production is an energy intensive procedure. In the blast furnace of steelmaking plants, coke serves multiple tasks as a reagent of chemical reductant, furnace burden support, and as a fuel. Coke is an expensive material and, quantitatively, is the largest component integrated into the blast furnace; therefore, the energy efficiency of steelmaking plants significantly depend on the quality of coke [2–7]. Properties of coal, as parent of coke, play an essential role in the metallurgical performance of coke (coke quality), and thus the energy consumption in the blast furnace. Feeding high quality coke to the blast furnace will result in lower coke consumption, higher steel productivity, and lower hot metal cost. Fundamentally, coal rank parameters (volatile matter, moisture, carbon, etc.) and petrographic composition (macerals) of coal are independent parameters that control the quality of coke [2,4,5,8,9]. Coking quality of coal can be determined by dilatation, fluidity, and Free Swelling Index (FSI) of coke samples [4,5].

In the United States, cokeability of coal is determined by FSI test (ASTM D720) [10]. FSI test can provide information regarding the caking ability of coal samples. In this test, 1 g of a fresh grind coal sample (−250 μm) in a standard sized silica crucible is heated approximately at 820 ± 5 °C for 2.5 min by either electric or gas furnace. Samples are cooled and the carbon residue (coke) is removed to assess the coal's swelling during heating. By comparing the size and shape of the coke button with a series of standard outlines and scaling a value from 0 to 9 at an interval of 0.5, the cokeability of samples is assigned. Based on the FSI standard (ASTM D720), swelling indexes of 0–2, 2–4, and 4–9 indicate weakly, medium, and strong caking ranges, respectively [10]. There are a few problems associated with the FSI measurement which lead to bias in the result of this test: providing the proper heating rate in the furnace, weathering of the sample (oxidation), and the size of samples (the amount of fine particles should be kept at a minimum)

* Corresponding author.
*E-mail address:* sos4552@gmail.com (S. Chehreh Chelgani).

http://dx.doi.org/10.1016/j.measurement.2016.07.070
0263-2241/

[11–13]. To overcome these challenges, statistical models based on empirical data of coal properties are applied to study coke quality more accurately, and to better control the parameters which impact energy consumption in the blast furnace.

According to these considerations, a few statistical models (regression and artificial neural networks (ANNs), Adaptive neuro-fuzzy inference systems (ANFIS), and genetic algorithm (GA)) have been used to evaluate the relationships between coal compositions (proximate, ultimate, and petrographic analyses) and its cokeability (FSI, coke reactivity index (CRI), and coke strength after reaction with carbon dioxide (CSR)) [7,11]. Moreover, soft computing methods (ANNs, ANFIS, GA, etc.) as intelligent techniques, are widely used in many areas in coal processing, such as prediction of Hardgrove Grindability Index (HGI) [14–17], Gross Calorific Value (GCV) [18,19], coal flotation [20–22], and desulfurization [23–26]. Generally these models have a conceptual limitation, since they can only determine relationships between input and output, but never give any insight into the interdependences among variables. In their calculation, the variances of the conditional distributions for the dependent variables are all equal and these models do not present variable importance measurement (VIM). VIM helps to select the best subset of predictors, this selection would lead to explain the model in the simplest way (the smallest model), remove redundant variables (decrease noise of predictions), and save time and energy by not measuring redundant predictors. These facts would be more critical in modeling of energy resources (especially coal and oil, due to their heterogeneous properties). Therefore, variable selection is necessary for the FSI modeling since involving all coal parameters as input in a model can potentially improve the correlation coefficient ($R^2$) of the model, but does not necessarily mean that the model can precisely describe the FSI [27,28].

Random forest (RF) models can reliably overcome these drawbacks. As a recent developed tree-based model, RF can identify nonlinear approximation of relationships among variables even for a high-dimensional database, and rank candidate predictors based on their inbuilt VIM. RF can be applied for prediction and classification of various problems including nominal, metric, and survival responses based on results of VIM (best predictors). RF models have several attractive features over other modeling methods, including: the most accurate and efficient nonparametric learning algorithms available, VIMs even in the presence of high levels of additive noise, highly accurate classifier, variables can be both continuous and categorical, automatic calculation of generalization errors (low-bias and low-variation in prediction), general resistance to overfitting, automatic handling of missing data, and a small number of tunable parameters [29–38].

Although there are numerous studies on using RF method as a new data mining tool, RF modeling and associated interpretations (via VIM) are not yet widely used in the engineering fields, especially within the energy sector. The main purpose of this article is to assess coke quality (FSI) based on various coal analyses (proximate, ultimate, various sulfur forms and ash oxides analysis) for a wide range of USA samples (3691 samples from 17 different states) by RF. Modeling and VIMs were carried out using the Random Forest "R" package.

## 2. Materials and methods

### 2.1. Database

Development of a realistic model for prediction of FSI requires a comprehensive dataset to cover a wide variety of coal properties. Such a model will be able to predict cokeability with a high degree of validity. In this investigation, the dataset used to study the proposed approaches was obtained from U.S. Geological Survey Coal Quality (COALQUAL) database, open file report 97-134 [39]. A total of 3961 set of coal samples including the proximate, ultimate, oxide, and FSI analyses in as received basis were used. Analyses were performed based on the standard ASTM test methods. The procedures of sampling and analytical chemical methods can be found on the following web address: http://energy.er.usgs.gov/products/databases/CoalQual/index.htm. The results of various analyses and their representative FSIs are presented in the supplementary database. The number of samples for different states is shown in Table 1.

### 2.2. Random forest

#### 2.2.1. Variable importance measurements (VIMs)

The aim of variable importance measurement is to identify the best subset between many variables to include in a model. VIMs help to better understand the fundamentals of a process, with best predictors. RF model can accurately predict a target, and cost can be saved by not measuring redundant predictors. In RF methods, the most efficient and advance VIM is the "permutation accuracy importance (PAI)" measure [40–45]. The PAI is to quantify the importance of predictors in function approximation. The PAI follows the rationale by determining the decrease accuracy of differences between the predicted value for a tree before and after random permutation for each predictor variable (i.e. with and without the association of each variable). In other words, PAI destroys the original association of a variable with the response. The average of differences over all trees determines the final importance score of a variable. Large value of the PAI implies the association between the predictor and the response is significant, values around zero indicate that the prediction accuracy would not increase with association of those variables and they have no value for predicting the response [29,42,46].

In summary, the computation of the PAI consists of the following steps: 1. Compute the out-of-bag (OOB) accuracy of a tree (the excluded examples construct called out-of-bag dataset). 2. Permute the predictor variable of interest in the OOB instances associated with a tree in the forest. 3. Recompute the OOB accuracy of the tree (destroying the information content of the covariate using the permuted variable). 4. Determine the error between the original and recomputed OOB accuracy. 5. Repeat step 1–4 for each tree, the average OOB difference over all trees is the indication of the overall importance score [31,42,47,48]. The PAI measurements have several advantages over other variable selection methods; it is unbiased, broadly applicable, and it considers the impact of each predictor individually as well as in multivariate interactions with other input variables [40,41,46,49]. For VIMs (variable selection), the reference implementation of PAI is available in the "R" software package (a free software package for statistical computing) which has been used in this investigation.

**Table 1**
Number of samples for the different states.

| State | *N* | State | *N* |
|---|---|---|---|
| Alabama | 733 | Ohio | 581 |
| Colorado | 96 | Oklahoma | 29 |
| Illinois | 16 | Pennsylvania | 354 |
| Indiana | 97 | Tennessee | 51 |
| Kansas | 21 | Utah | 66 |
| Kentucky | 798 | Virginia | 320 |
| Iowa | 53 | West Virginia | 366 |
| Missouri | 65 | Wyoming | 16 |
| New Mexico | 29 | | |