Article:

Morris, J., Daood, S. and Nimmo, W. orcid.org/0000-0001-5571-026X (2022) Machine learning prediction and analysis of commercial wood fuel blends used in a typical biomass power station. Fuel, 316. 123364. ISSN 0016-2361

https://doi.org/10.1016/j.fuel.2022.123364

Machine learning prediction and analysis of commercial wood fuel blends used in a typical biomass power station

Jonathan D. Morris[a], Syed Sheraz Daood[a,b,c]*, William Nimmo[a]

[a] *Energy Engineering Group, Energy 2050, Department of Mechanical Engineering, University of Sheffield, Sheffield, S3 7RD, United Kingdom*

[b] *Institute of Chemical Engineering and Technology, University of the Punjab, Quaid-e-Azam Campus, Lahore, Pakistan*

[c] *Energy Engineering Research and Development Centre, University of the Punjab, Quaid-e-Azam Campus, Lahore, Pakistan*

** Corresponding author at: Institute of Chemical Engineering and Technology, University of the Punjab, Quaid-e-Azam Campus, Lahore, Pakistan*

E-mail addresses: sdaood.icet@pu.edu.pk (S.S. Daood); w.nimmo@sheffield.ac.uk (W. Nimmo)

Abstract

Biomass can be utilised as a near carbon neutral fuel for power generation. Biomass may come in a variety of forms, such as woods or agricultural wastes, with highly variable compositions. It is well known that biomass ash content and composition can introduce severe operational challenges to power stations, such as slagging and corrosion. Machine learning approaches have been successfully applied in a variety of contexts to generate predictive models and identify relationships in large data sets. However, such approaches have not previously been applied to biomass trace element or ash content data, in part due to limited large data set availability. In this work, 5 years of fuel blend analysis data (3500+

data sets) was analysed from a 35MW$_e$ biomass power station burning a 60/40% blend of virgin wood and recycled (waste) wood. Variation to ash content and key trace elements (K, Na, Pb, Zn, Cl) was analysed and compared versus a literature average benchmark. Identification of underlying relationships between these key components and others was attempted with principal component analysis and random forest regression machine learning. Ash and chlorine exceeded the literature average benchmark over many time periods and would have a negative impact on boiler operation. Potassium and sodium were only above literature average levels intermittently. No significant underlying relationships for the key components could be identified with principal component analysis or random forest regression, nor could an accurate predictive model be created, though some minor trends were noted. This is likely due to the high degree of heterogeneity seen in the fuel data, as it is a blend of virgin and recycled wood. It is suggested that future studies applying machine learning methods in this context either use singular fuel data sets, or that additional information is recorded within the blended data set when analysis fuel composition (e.g. fuel sources, suppliers, blend ratio). Suggestions were also made regarding improvements to waste wood sampling approaches.

Keywords

Biomass, waste wood, ash, random forest

Nomenclature

ANN	Artificial neural network
CHNS	Carbon Hydrogen Nitrogen Sulphur
GCV	Gross calorific value
MAE	Mean absolute error

MAPE	Mean absolute percentage error
MSE	Mean square error
OOB	Out-of-bag
PC	Principal component
PCA	Principal component analysis
PVC	Poly-vinyl chloride
R^2	Coefficient of determination
RF	Random Forest
RMSE	Root mean squared error
UKAS	United Kingdom Accreditation Service
$\hat{y}$	Predicted value (model error analysis)
y	Actual value (model error analysis)
n	Number of predicted points (model error analysis)
s	Score value (permutation importance)

1. Introduction

Increasing awareness of the effects of anthropogenic climate change has led to the rapid uptake of low carbon options for power generation. One such near carbon-neutral fuel option is biomass, which encompasses materials such as wood or agricultural wastes [1]. Biomass fuels are known to bring many challenges for combustion due to their heterogeneous nature [2, 3], such as slagging, fouling, corrosion, and in fluidized bed boilers, bed agglomeration. Key fuel parameters for combustion, such as ash and moisture content, can vary substantially by fuel type, fuel harvesting location or season, or fuel pre-processing methods (e.g. washing [4]). Both experimental and theoretical studies of biomass fuel compositions and challenges have been performed, to better understand their fundamental characteristics.

1.1 Biomass Experimental Studies

Biomass power station operators frequently prefer virgin wood fuels which generally have more desirable and consistent compositions (e.g. lower ash contents). The higher cost of virgin wood fuels has led some operators to utilise these fuels in a blend with cheaper, lower quality, waste wood. As is well known in the literature, use of waste wood comes with many issues for boiler operation [5]. Unavoidable extraneous contaminants (e.g. plastics, paints, metals) contribute high levels of chlorine, zinc, and lead in the feedstock, which drive corrosion within the boiler [6, 7, 8]. These extraneous contaminants are frequently under-reported in fuel composition analysis, with accurate sampling of recycled wood challenging due to its heterogeneity [9]. A further concern with recycled wood is higher levels of nitrogen, introduced in the manufacture and preparation of the wood for its original use (e.g. in glues) [10], which can contribute to higher NO_x emissions. Having reliable predictive models for fuel quality would be of significant value to operators. At present, it often takes days for fuel blend analysis to be performed, despite sampling occurring at the point of fuel entry into the boiler where it is immediately burned. Identifying reliable correlations between fuel contaminants may allow, for example, to monitor one trace element via a boiler measurement probe and use this to determine levels of correlating trace elements.

1.2 Biomass Fuel Modelling

Various modelling approaches have been applied in the study of biomass fuels. Thermochemical modelling software packages, such as FactSage [11], have been applied in the study of biomass ash melting behaviours, to determine if a given fuel is likely to cause ash related issues (e.g. slagging) during combustion. These approaches use fundamental thermochemical parameters for unary, binary, ternary, and quaternary chemical phase systems, in combination with Gibbs free energy minimization, to predict compound and phase formations at a specified set of conditions. The recent work of Lindberg, et al. [12] reviewed the current status of the thermochemical databases that underpin thermochemical modelling approaches, and found key ternary systems (e.g. K_2O-CaO-SiO_2) to lack fundamental experimental data, which would thus negatively impact the accuracy of thermochemical models using these databases.

Other mathematical studies of biomass fuels have examined correlations between oxides within ash and differences between biomasses (e.g. woody fuels versus contaminated biomass wastes), as was performed by Vassilev, et al. [3]. This work highlighted for example the positive correlations in natural biomass with Cl, Ti, Si, S, Na and P that would arise in ash content. Further examples of mathematical modelling include the work of Edo, et al. [5] where a 9-year waste wood fuel data set from a Swedish power station was analysed, one of the few available works that has studied a large commercial biomass fuel data set. This study applied a handful of mathematical approaches, such as principal component analysis, to the fuel data set to attempt to identify underlying groupings or correlations. Edo, et al. [5] noted a correlation between waste wood particle size and composition, e.g. sieving to reduce biomass fines had also reduced chlorine and lead content.

In recent years, machine learning methods have rapidly captured interest as method to better understand large data sets. Random forest is an ensemble machine learning algorithm [13, 14, 15] that has been widely applied across academia and industry. For example, a search for the term “random forest” in the citation database Scopus [16] returns over 48,000 results (as of 2021), with the majority published since 2015. With random forest, many separate decision trees are created in model training, with each of these trees making a prediction. The aggregate result of these trees is then used to make the final prediction. In the context of a regression study, the prediction results of all decision trees are averaged to create the final prediction.

Few prior works have applied random forest in the context of biomass fuel quality studies [17, 18]. Ge, et al. [17] trialled random forest and other decision tree approaches to identify fuel types based upon flame spectra data. Four different biomass fuels were used in testing. A total of 4000 data points were collected, with seven data features used for prediction in the tree models. After tuning of the random forest model, the average identification success rate after ten trials with the model was 98.7%. Elmaz, et al. [18] applied random forest classification to proximate fuel analysis data, to classify fuels into one of four types: coal, wood, agricultural residue, or manufactured biomass. Other models were used in addition to random forest, on close to 600 fuel data sets taken from the Phyllis2 [19] database. With random forest, a classification accuracy of about 90% was seen. No prior literature studies appear to have applied random forest for the prediction of key fuel compositional parameters or trace element contents. As can be imagined, a variety of other machine learning and artificial intelligence methods have been applied more broadly in the context of biomass fuel usage. For example, the recent work of Li, et al. [20] applied artificial neural networks for the prediction of potassium content based on flame spectroscopy; potassium playing a key role in ash melting and corrosion phenomena. The best neural network applied, a deep recurrent neural network, had a low relative error of 6.34% which is a good endorsement of the viability of neural networks when using flame spectroscopy data. A further recent example of neural network application in the context of biomass fuel prediction is the work of Sakiewicz, et al. [21] who used this approach to predict ash fusion melting temperatures. The accuracy achieved was however relatively weak, with R^2 values between 0.615-0.756.

1.3 Study Background and Objectives

In the present work, a 5-year fuel data set from a $35MW_e$ biomass fired power station was analysed. The unit uses a bubbling fluidized bed boiler and burns a blend of virgin wood and recycled (waste) wood. The target fuel blend is 60% virgin wood and 40% recycled wood on a thermal basis, though recycled wood content may vary from 25-45% dependent fuel quality. All fuel is sourced from the UK. Virgin wood is comprised of a variety of sources (logs, chips, sawmill residues) from several suppliers. The recycled wood is also sourced from around the UK from several suppliers and is typically in the form of building/demolition waste wood. The fuel data under analysis was the “blend-to-boiler” analysis. This is the analysis of samples of the blended fuel mixture taken from the screw feeder, just prior to entry into the boiler. This therefore presents a challenging scenario for data analysis and machine learning approaches, as this is a highly heterogeneous data set with various factors changing simultaneously (e.g. blend ratio, fuel source/supplier, trace element levels, etc.). In general, very few studies of large industrial biomass fuel data sets exist in literature. As previously mentioned, the most notable example is that of Edo, et al. [5] where a 9-year waste wood fuel data set from a Swedish power station was analysed, however this did not apply machine learning methods.

The objectives of the study were to:

- Analyse the levels of key components (ash, K, Na, Cl, Pb, Zn), including changes over time.
- Perform random forest regression on the blended fuel data set, to see if this method can accurately predict the aforementioned key ash and trace element levels and identify previously unknown relationships.

2. Methods

2.1 Data Source

The primary data source for this study was the fuel blend-to-boiler analysis data covering the period 2014-2019 for a 35MW$_e$ biomass fired power station burning a target blend of 60% virgin wood and 40% recycled (waste) wood on a thermal basis. This data consists of a full ultimate analysis and fuel trace element analysis of fuel blend samples, with the analysis performed by the UKAS accredited power station lab team. A full list of parameters that are captured in this analysis are listed in Table 1. Samples are taken twice a day from a sampling point just prior to the fuel entering the boiler. Samples are subject to the following analysis:

- CHNS analysis.
- Ash content (with muffle furnace).
- Bomb calorimetry, with residues analysed for halide content with ion chromatography.
- Metals content analysis via inductively coupled plasma mass spectrometry.

Table 1: *List of fuel analysis parameters subject to measurement.*

No.	Feature	Units
1	Gross calorific value	GJ/te
2	Moisture	% a.r.
3	Carbon	% (dry)
4	Hydrogen	% (dry)
5	Nitrogen	% (dry)

No.	Feature	Units
6	Sulphur	% (dry)
7	Fluorine	mg/kg dry (fuel mass)
8	Ash	% (dry)
9	Chlorine	mg/kg dry (fuel mass)
10	Arsenic	mg/kg dry (fuel mass)
11	Cadmium	mg/kg dry (fuel mass)
12	Chromium	mg/kg dry (fuel mass)
13	Copper	mg/kg dry (fuel mass)
14	Mercury	mg/kg dry (fuel mass)
15	Nickel	mg/kg dry (fuel mass)
16	Lead	mg/kg dry (fuel mass)
17	Vanadium	mg/kg dry (fuel mass)
18	Zinc	mg/kg dry (fuel mass)
19	Antimony	mg/kg dry (fuel mass)
20	Cobalt	mg/kg dry (fuel mass)
21	Manganese	mg/kg dry (fuel mass)
22	Thallium	mg/kg dry (fuel mass)
23	Tin	mg/kg dry (fuel mass)
24	Aluminium	mg/kg dry (fuel mass)
25	Sodium	mg/kg dry (fuel mass)
26	Potassium	mg/kg dry (fuel mass)
27	Particle Size: >80mm	mm
28	Particle Size: 50-80mm	mm
29	Particle Size: 5-50mm	mm
30	Particle Size: <5mm	mm

2.2 Fuel Quality Variation Over Time

Six key fuel components in the fuel blend data (Ash, Cl, Pb, Zn, K, Na) were analysed for variation over time, with percentile, mean, median, and standard deviation values tabulated. Outliers were removed for the creation of quarterly averages which were then graphed, and this outlier removal process is discussed in subsequent section 2.3.

Quarterly average values were evaluated against calculated literature averages for the blend, shown in Table 2. For virgin wood, an average of softwood data (fire, spruce, and pine) was taken from the Phyllis2 database [19], which collates biomass fuel analysis data from lab and literature sources. For each of the six key components, between 12-105 samples were used to calculate the average depending on the available data for each in Phyllis2. For recycled wood, data from the work of Edo, et al. [5] was used, which analysed 500 waste wood samples over a 9 year period from a biomass power station. Averages for virgin and recycled wood were then weighted to create a literature average for the blend under analysis. The 60/40% (energy basis) virgin wood/recycled wood blend was calculated as a 72/28% blend on a mass basis, using representative target moisture values.

Table 2: *Literature average values for virgin wood, recycled wood, and the blend, that were used as an evaluation benchmark for the fuel blend data studied here. Recycled wood average from Edo, et al. [5]. Virgin wood average from softwood data in the Phyllis2 database [19].*

Component	**Unit**	**Recycled Wood**	**Virgin Wood**	**Blend Average** (60/40% Virgin Wood/Recycled Wood Energy Basis; 72/28% Virgin Wood/ Recycled Wood Mass Basis)
Ash	wt.% (dry)	4.39	0.83	1.83
Chlorine (Cl)	mg/kg (dry)	1,000.0	350.1	532
Potassium (K)	mg/kg (dry)	1,032.0	1,209.8	1160
Sodium (Na)	mg/kg (dry)	967.0	107.2	348
Lead (Pb)	mg/kg (dry)	81.0	1.8	24
Zinc (Zn)	mg/kg (dry)	450.0	38.7	154
K + Na	mg/kg (dry)	1,999.0	1,317.0	1508

2.3 Principal Component Analysis

Principal component analysis (PCA) is a widely used dimensional reduction technique that reduces a data set comprising of many variables to a handful of principal components (PCs) [22]. These PCs can then be used as a new set of variables for further analysis, or if meaningful, may be interpreted to describe a phenomenon or grouping in the data [22, pp. 63-64]. For example, rainfall readings at meteorological stations A, B, and C that are all in city X can instead be reduced to “PC1” which describes rainfall in city X.

PCA was applied to the six key fuel components (Ash, Cl, Pb, Zn, K, Na), plus fuel particle size, to determine if any underlying relationships existed between these parameters. This was performed via a custom script written in MATLAB R2019b [23]. The MATLAB script for PCA is presented in the Supplementary Data.

Table 3: *Maximum and minimum values for the fuel data set subject to PCA analysis, before and after the removal of outliers.*

Variable	**Original data set**		**Refined data set**	
	Minimum	**Maximum**	**Minimum**	**Maximum**
Ash (wt.%)	0.15	26.42	0.5	10
Cl (mg/kg)	0.1	20839	100	2500
Pb (mg/kg)	0.111	2473	10	300
Zn (mg/kg)	1.434	8652	15	500
Na (mg/kg)	0	3112.5	10	1300
K (mg/kg)	7.754	14866	200	2500
Particle size: >80mm (%)	0	82.83	*0**	*40.05**
Particle size: 50-80mm (%)	0	38.37	*0**	*38.37**
Particle size: 5-50mm (%)	7.55	98.60	*7.55**	*97.97**
Particle size: <5mm (%)	1.34	73.34	*2.5**	*73.45**

** A maximum/minimum cut-off was not used for particle size. These were the respective maximum and minimum values remaining after refining data for the other six variables.*

Outliers were removed from this seven-variable group, leaving 2786 data sets from a starting amount of 3823. The reduction bounds used for outlier removal are shown in Table 3. Where a data point was removed for being an outlier, the entire data set for that time stamp was removed. Weighted average particle size was calculated by taking the mid-point of the particle size grades to calculate a weighted average. As part of the MATLAB script, data was normalized on a 0-1 scale due to the use of different units between the variables. This was to prevent unintentional component bias in PCA when working across units and scales.

2.4 Random Forest Regression

2.4.1 Model Configuration, Creation, and Validation

Random forest regression was applied to predict values of the six key fuel components, Ash, Cl, Pb, Zn, K, Na. The aim was that if a successful predictive model were generated, it would in turn have revealed the necessary underlying correlations to make such a prediction. The script implementing random forest was written in Python 3.8.5, using the Anaconda Python distribution and the Spyder integrated development environment. The source code is available in the Supplementary Data.

The fuel data set had a total of 30 data features (i.e. measurement variables, shown in Table 1) with 2786 data sets, after removal of empty rows and large outliers. This “cleaned” data set had the same data cleaning/outlier removal criteria applied as for the principal component analysis (cleaned data bounds shown in Table 3), with the difference being for the random forest data set all 30 data features remained in the data set as opposed to only seven parameters in the PCA analysis. Six different fuel parameters were selected as prediction targets (“y-values”): Ash content, Cl, K, Na, Pb, Zn. In each modelling case, the remaining 29 data features were thus the input x-variables.

The random forest regression model used near-default model hyperparameters, as tabulated in Table 4. Model hyperparameter tuning was attempted, however, this tuning process was not found to result in a tangible improvement to prediction accuracy, in addition to being computationally intensive.

To ensure the correct function of the model, a synthetic fuel data set was created containing random and correlated data (‘y = mx’ correlation). The model accurately identified this correlation. With the entirely random data set, whilst the model could reasonably “fit” during training, it lost all accuracy when used with the unseen random testing data set, as would be expected. These exercises are described in the Supplementary Data.

Table 4: *Random Forest hyperparameters used, as well as the test/train split for the input data.*

Parameter	Definition	Setting
n_estimators	Number of trees in the forest.	100*
max_features	Number of features selected at each splitting node as basis on which to split data.	10[a]
max_depth	Maximum levels of growth of each tree.	None*
min_samples_split	Minimum number of samples required to split a node.	2*
min_samples_leaf	Minimum number of samples required at each leaf after a split.	1*
bootstrap	Whether to bootstrap input data for each tree, i.e. sample with replacement.	True*
Test / Train Split	Proportion of data assigned to model training and testing.	70% training 30% testing[b]

** Denotes the value is the default value for the random forest regressor function.*

[a] *Was selected as this is near-equal to p/3, where p is the number of input data features (29 in this case) used for the prediction of the target feature. A ratio of p/3 is a common 'rule of thumb' for random forest regression models* [15, p. 592].

[b] *Initial exploratory tests with lower or higher allocations to model training did not give an appreciable improvement to predictions.*

2.4.2 Evaluation of Model Predictions

For each prediction case, two different results plots were created. Predicted y-value versus the real y-value for the testing data set was plotted to visualise the overall prediction quality. Permutation importance was also plotted. Permutation importance is a common measure of the importance of a data feature (i.e. an input x-variable) for prediction accuracy. Permutation importance works as follows: a baseline accuracy score for the model, using either training or testing data, is first acquired. Then, using the testing data set, each feature (column) of input data is randomly shuffled, and the accuracy score then calculated when using this shuffled column. The difference between the baseline and randomly shuffled accuracy scores is then evaluated, to determine the permutation importance of each variable to the prediction. Shuffling a data feature that does correlate to the predicted output should naturally incur a substantial penalty to prediction accuracy, and hence have a high permutation importance.

The formula for permutation importance in Scikit-Learn is shown in Equation **2.1** [24]. In this, 'i' is the importance value for the feature under evaluation (e.g. analysing the importance of chlorine data for the prediction of ash content), 's' is the score, and 'k' is the number of times the feature is to be randomly shuffled. This would then be applied for all input data features used in the prediction model. The coefficient of determination, R^2, was used as the scoring metric 's'.

$$i_j = s - \frac{1}{K}\sum_{k=1}^{K} s_{k,j}$$

Equation 2.1

Several error measures were used to evaluate prediction performance of the model: mean absolute percentage error (MAPE), mean absolute error (MAE), mean square error (MSE), root mean squared error (RMSE), coefficient of determination (R^2), and the out-of-bag (OOB) error.

MAPE is defined in Equation **2.2**, where ‘n’ is the number of points predicted by the model, ‘y’ is the actual value and ‘ŷ’ is the predicted value. A lower value of MAPE is better.

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

Equation 2.2

MAE is defined in Equation **2.3**, where ‘n’ is the number of predicted points, ‘ŷ’ is the predicted value and ‘y’ is the actual value. A lower value of MAE is better.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i|$$

Equation 2.3

MSE is defined in Equation **2.4** [25]. Again, ‘n’ is the number of predicted points, ‘y’ is the actual value, and ‘ŷ’ is the predicted value. The RMSE is the square root of the result of the MSE therefore is not explicitly shown here. Lower values of MSE and RMSE are better.

$$MSE = \frac{1}{n}\sum_{i=0}^{n-1}(y_i - \hat{y}_i)^2$$

Equation 2.4

The coefficient of determination, R^2, is defined in Equation **2.5** [25]. In this equation, ‘y’ is the real value, ‘ŷ’ is the predicted value, and ‘ȳ’ is the mean of the actual data. Values of R^2 closer to 1 indicate a better fit of the predicted versus actual values.

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Equation 2.5

The out-of-bag (OOB) error is further measure of accuracy. Random forest uses bootstrapping of sample data for each decision tree. This means that not all training data is used for training each tree. The OOB score is an evaluation of the prediction error when using data samples that did not form part of the bootstrapped data sample as the inputs to the model. The OOB error is simply (Equation **2.6**):

$$OOB_{error} = 1 - OOB_{score}$$

Equation 2.6

The OOB score can be evaluated on many different metrics (e.g. MSE, RMSE, R^2, etc.).

3. Results and Discussion

3.1 Analysis of Blend-to-Boiler Composition

An analysis of key trace elements and ash content in the fuel data is given in Figure 1, presented on a quarterly average basis. Significant outliers have been removed from the data set, as discussed in sections 2.2 and 2.3. Also shown on Figure 1 are the average literature benchmark values as calculated in section 2.2, which are labelled “Lit. Avg.”. The percentile values of these six key components are tabulated in Table 5 to provide further detail, with outliers included in this tabulated data. The mean, median and standard deviation for each of the six key components, with and without outliers, are listed in Table 6.

There are a few trends evident in Figure 1. There is a reduced amount of K+Na and ash in the 2016-Q2 to 2017-Q4 period, possibly in part due to a change in the fuel source or fuel blend make-up. For example, the quantity of brash wood or bark in the virgin wood blend may have been reduced, with these seen in literature to have higher amounts of ash and K+Na than stem wood [26, 3]. However, given the 2wt.% decline in ash content, it is also possible that these ash and K+Na behaviours are coincidental, given that waste wood typically has significantly more ash content than virgin wood [3], and there was a near halving of ash content over the 2016-Q2 to 2017-Q4 period.

On ash content in Figure 1, it can be seen clearly that ash content is almost always above the literature average value of 1.83wt.%, through in some time periods, e.g. 2016-Q4 to 2017-Q4, it is close to this average. This is likewise evident in Table 5, where the top 80% of samples all exceed the literature average of 1.83wt.%. It should be noted that this is an evaluation against average literature values. In other words, the maximum allowable value, or boiler limit, is likely to be higher than this literature average. However, it does show that there is clear room for reduction in ash content, and this was achieved in the 2016-Q4 to 2017-Q4 period. Whilst detailed and reliable analysis data for virgin wood and recycled wood batches was not available, it is likely that the ash predominantly came from the recycled wood fraction, with ash levels in waste/recycled wood shown to be over four times higher than those in virgin wood as per the literature averages calculated in section 2.2. A reduction in fuel ash content would provide many benefits, as it would reduce quantities of problematic ash components, reduce wear on ash handling equipment, and minimise corrosion, slagging, and agglomeration issues [27]. An important factor when discussing ash content is that some extraneous contaminants from the recycled wood fraction (e.g. nails, plastics, etc.) that may bypass screening measures to some extent are not subject to analysis and would not be included in this ash content figure. It is difficult to estimate the potential increase in total ash content arising from these components. Others have found these to account for 1.1wt.% of waste wood by weight [5]. It would be recommended to begin accounting for these extraneous contaminants by mass as an additional data set, as they can contribute to quantities of other components, e.g. zinc from galvanised metals [28], or chlorine from plastics [9]. Similar studies have been performed on industrial boiler fuel sources by others and were seen to provide useful insight into fuel quality [5].

With regards to the combined K+Na content in Figure 1, this value was mostly below the literature benchmark average of 1508mg/kg (dry). A few quarters over the five-year period showed averages of around 1700-1900mg/kg. In Table 5, whilst potassium and sodium are presented on an individual basis, it is clear that outside of the 100th percentile, sodium is always below 350mg/kg, whereas for potassium upwards of the 80th percentile of samples is in excess of the combined K+Na literature average of 1508mg/kg. This is integral to avoiding unplanned outages because both alkali metals drive slagging, agglomeration, and corrosion mechanisms [27]. There is little that could be done to further reduce or control quantities of potassium, as it is a nutrient for forest growth [29]. The only options available would be to switch to a lower potassium wood source, or to water wash the biomass [4], both of which would incur additional cost and are only applicable to the virgin wood fraction.

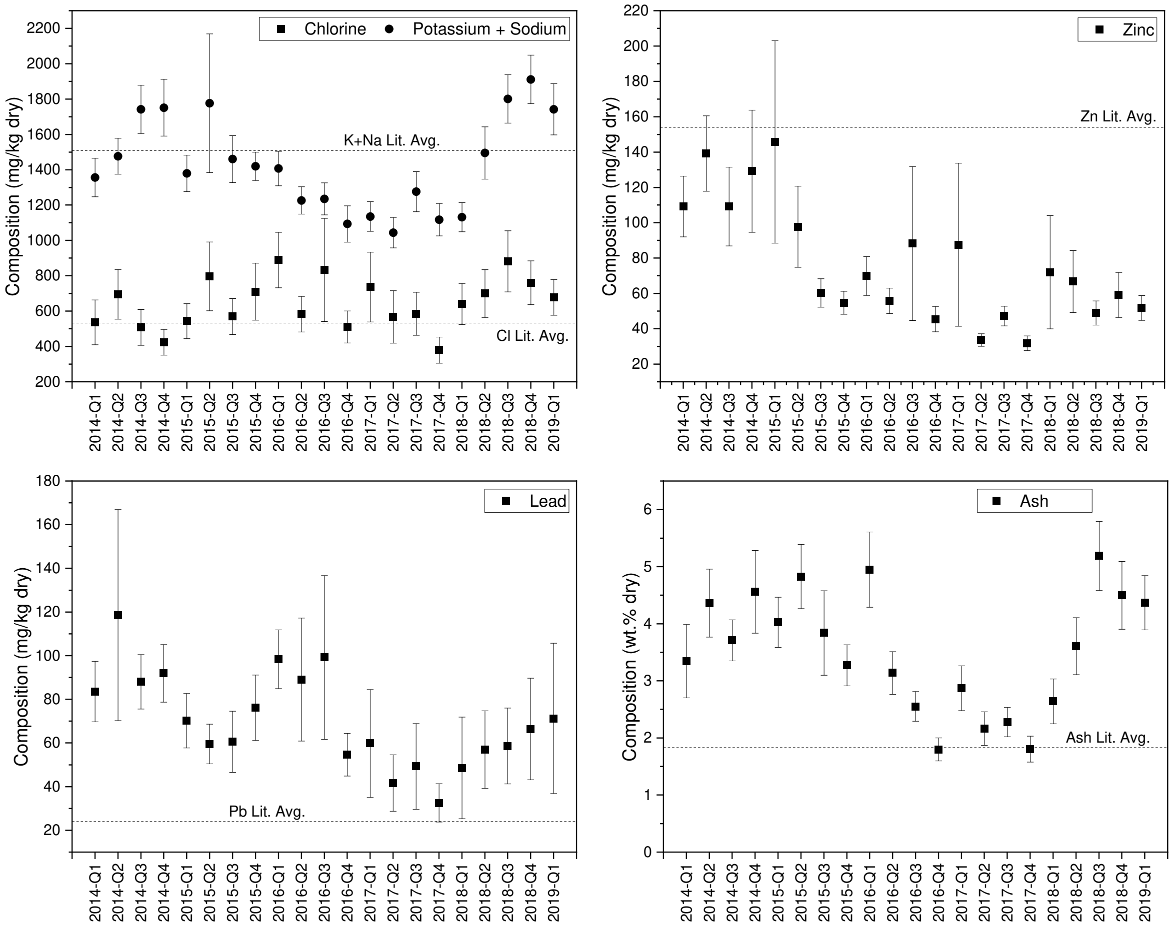

***Figure 1:** Quarterly average of key fuel components in the fuel blend with 95% confidence intervals shown. Also shown are the literature average benchmark values (labelled "Lit. Avg.") for each component, as calculated in section 2.2.*

Both lead and zinc content are primarily derived from the recycled wood fraction of the fuel blend. These components cause corrosion through the formation of lead and zinc chlorides [27, 28], which can condense onto superheater surfaces, or can worsen ash melting issues in the bed. It can be seen from Figure 1 that zinc is generally substantially below the 154mg/kg literature average value, and that levels of zinc have decline significantly since the 2014-Q1 to 2015-Q2 period. There is however some substantial variance to zinc content exhibited, as shown by the large 95% confidence intervals in many quarters. This variance is also evident when comparing the mean and median values for either the original or refined data sets in Table 6, where the median is around 55-60mg/kg in either case, versus a mean upwards of 100mg/kg in the original data set with outliers. Around 20 outlier data points were removed from the zinc data set for Figure 1 as they had values upwards of 800mg/kg, into the thousands, which were removed to prevent distortion of the averages. Table 5 has the outliers included for Zinc, and it is evident that the top 10% of samples have zinc levels higher than the 154mg/kg literature average. Samples with very high zinc contents were likely the result of a high level of extraneous metals (e.g. galvanised coatings) in a sample. This highlights again the both the importance and difficulty of accurate sampling, particularly with waste wood, and the need for some form of accounting of extraneous contaminants that might not be subject to analysis but may find their way into the boiler at some level. With regards to lead content, similar to ash content it is apparent that levels are almost always in excess of the literature average of 24mg/kg. Once again, it must be noted that this is an average rather than a maximum, but it does suggest that these levels could be reduced. From Table 5, there were some outlier samples excluded in Figure 1 for lead, with the 100th percentile being 1274.7mg/kg, likely due to high extraneous lead contents in the sample.

Table 5: *Values of key fuel components at different percentiles of the normal distribution for the fuel data set. For reference, literature average values as calculated in section 2.2: Cl 532mg/kg, Pb 24mg/kg, Zn 154mg/kg, K+Na 1508mg/kg, Ash 1.83wt.%.*

Percentile of Normal Distribution	**Chlorine (mg/kg dry)**	**Lead (mg/kg dry)**	**Zinc (mg/kg dry)**	**Sodium (mg/kg dry)**	**Potassium (mg/kg dry)**	**Ash (wt.%)**
10%	297.2	27.3	32.0	111.1	801.5	1.7
20%	369.3	35.9	37.9	136.3	930.3	2.2
30%	428.8	43.2	43.5	153.2	1017.0	2.6
40%	484.1	51.5	49.6	175.5	1102.1	3.0
50%	545.3	59.9	58.2	194.3	1186.5	3.4
60%	621.4	68.3	68.2	214.2	1273.5	3.7
70%	707.6	79.4	83.3	233.1	1381.1	4.2
80%	846.6	95.6	108.1	257.4	1512.9	4.8
85%	974.0	107.3	127.0	277.7	1612.5	5.1
90%	1120.0	120.4	157.5	302.0	1743.5	5.7
95%	1404.0	154.3	218.8	334.0	1924.5	6.6
99%	1561.5	172.6	298.5	347.1	2052.0	7.0
100%	14165.5	1274.7	6879.5	1670.4	8124.9	16.0

***Table 6:** Table summarizing the mean, median and standard deviation for the refined data set without outliers (shown in Figure 1), and the original data set (with outliers) used for the percentile data in Table 1. Maximum and minimum values for these data sets are listed in Table 3 for reference.*

Component	**Original Data (with outliers)**			**Refined Data (without outliers)**		
	Mean	**Median**	**Standard Deviation**	**Mean**	**Median**	**Standard Deviation**
Ash (wt.% dry)	3.60	3.30	1.89	3.52	3.29	1.63
Cl (mg/kg dry)	664.2	534.3	632.5	625.2	534.0	357.6
Pb (mg/kg dry)	72.7	57.2	85.8	65.8	57.1	40.1
Zn (mg/kg dry)	112.8	56.9	353.6	74.8	56.2	59.3
Na (mg/kg dry)	202.4	192.8	107.0	202.8	193.4	92.7
K (mg/kg dry)	1234.3	1181.2	405.2	1226.5	1182.9	370.3

Chlorine is one of the most problematic components within biomass fuels due to its reaction with potassium to form KCl, driving alkali chloride induced corrosion [27, 30]. It can be seen clearly in Figure 1 for more than half of the quarters over the five year period, the average chlorine content exceeds the 532mg/kg literature average benchmark for this blend. This is also evident in Table 5, where the 50th percentile value is 545.3mg/kg, which is in excess of the literature average. In some quarters, e.g. 2016-Q1, the chlorine content is almost 900mg/kg, or around 70% above the literature average chlorine content. This chlorine content is likely to be primarily derived from the recycled wood fraction and its contaminants such as plastics. It is also worth noting that the real chlorine content entering the boiler may be higher still, due to the presence of plastics (e.g. poly-vinyl chloride, PVC) that are not fully accounted for during fuel blend sampling and analysis, but may still pass material screening to some extent and are fed into the boiler [9]. One way to gauge the real levels of chlorine inside the boiler environment, and equate them to those seen in fuel blend sampling, would be to use an in-situ alkali chloride monitor (IACM) or similar to measure alkali chloride levels in the boiler freeboard or back-pass [31], but such equipment was not installed on the boiler.

Studies of recycled wood fuels in the literature have reported high chlorine contents. For example, a nine-year study of a Swedish recycled wood combustion plant by Edo, et al. [5]

reported chlorine contents between 0.07-0.13wt.%, with a mean of about 0.1wt.% as noted in Table 2. Huron, et al. [9] tested a variety of types of waste wood (e.g. pallets, particle board, mixed sources). Four out of six of their representative waste wood blends had a chlorine content below 0.06wt.%, with the remaining two at 0.075wt.% and 0.12wt.% respectively. They observed that coated particle board was by far the most problematic waste wood component, with a chlorine content of 0.16wt.%. The higher chlorine content of waste wood should be offset by the virgin wood in the blend, and the fuel data under analysis here is only 40% recycled wood on an energy basis with the remainder being virgin wood. As seen from the literature average in section 2.2, virgin softwood as was used in this power station typically has an average chlorine content of 350.1mg/kg. Ad-hoc sample data for batches of virgin wood used in the power station were in the 200-400mg/kg range, which is in line with the literature average for virgin wood of 350.1mg/kg. This would suggest that the recycled wood used here has a particularly high chlorine content, hence the overall chlorine levels being higher than the calculated literature average for the blend of 532mg/kg. This may be due to higher-than-average levels of certain types of chlorine-rich materials in the recycled wood fraction, e.g. coated particleboard, plastics.

Fuel nitrogen content is the primary source of NO_x emissions in fluidized bed boilers, due to their lower operating temperatures versus traditional pulverized fuel boilers [32, 33]. It is known that different types of wood releasing different proportions of nitrogen content as NO_x [34, 10], and NO_x emissions are an integral part of the industrial emissions directive [35]. Fuel nitrogen content was separately analysed along with boiler emissions data, and it was found NO_x emissions were not at levels of concern with the current fuel blend.

3.2 Principal Component Analysis

The variance and cumulative variance for each of the seven PCs is presented in the form of a pareto chart in Figure 2 (the seven input variables being ash, K, Na, Cl, Pb, Zn, weighted average particle size). As can be seen, four of the seven PCs are required to account for 75% of the total variance. For a data set that is relatively small in dimensionality with only seven variables, this is quite poor. If two or three PCs accounted for most of the data set variance, this would suggest some underlying correlation(s). This suggests that either the data is too heterogeneous, or that there are no underlying trends between the seven input variables.

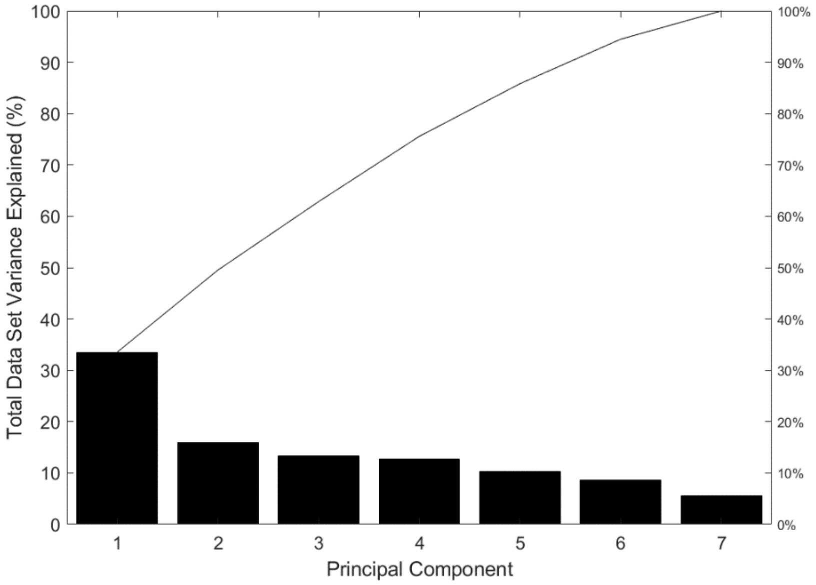

Figure 2: *Data set variance explained by each of the principal components, on both an individual principal component variance basis (bar) and cumulative variance basis (line).*

The coefficients for each principal component are listed in Table 7, with the data in Table 7 also represented graphically in Figure 3. The determination as to whether a principal component correlates strongly to a variable is largely subjective. Here, a value greater than 0.35, or less than -0.35, was assumed to be a notable correlation to a PC. These values have been highlighted in bold in Table 7. PC1 correlates moderately with ash content, and slightly less so with all other fuel measurement data except for chlorine and particle size. PC2-4 are moderately to strongly correlated to both Cl and particle size, in addition to 1-2 other fuel components. Aside from this non-correlation between ash content and Cl/particle size, there does not appear to be any major patterns across PC1-4, despite accounting for 75% of the total data set variance. The lack of correlation between ash, Cl and particle size may be due to fuel sampling. As discussed, a significant proportion of Cl content is likely to be extraneous, arising from plastics, therefore may not always be subject to analysis. Particle size is measured across large bounds (<5mm, 5-50mm, 50-80mm, 80mm+), and this lack of granularity may hide some trends. As noted in section 1.2, the closest comparable study of a large industrial biomass fuel set was by Edo, et al. [5], which examined a waste wood fuel only. Edo, et al. [5] applied PCA and observed a correlation between finer particle size of waste wood, and increases to chlorine and lead content, hence PCA is an appropriate and established method in a fuel data analysis context. The inability to identify distinct correlations in the present study may be due to the poor fuel particle size granularity, or lack of inclusion of extraneous materials in analysis, as previously mentioned. Alternatively, it may be because a blended wood fuel data set has been examined, instead of just a waste wood data set as was the case in the work of Edo, et al. [5].

It is also of interest that zinc is moderately well correlated within PC1 and PC2, as it has previously been noted (section 3.1) that most zinc is likely to arise from extraneous components (e.g. galvanised metals). This may suggest that more of these galvanised coatings end up within fuel samples for analysis than might otherwise be expected. However, given the generally weak correlations seen in this analysis, it is not possible to reach a definitive conclusion.

Table 7: *Coefficients for each principal component/variable pair.*

Variable	**PC1**	**PC2**	**PC3**	**PC4**	**PC5**	**PC6**	**PC7**
Ash	**0.51**	0.27	-0.25	-0.01	-0.26	0.06	**0.74**
Cl	0.19	**0.54**	**0.59**	**-0.48**	0.09	-0.27	-0.09
Pb	**0.41**	-0.32	**0.39**	-0.11	-0.10	**0.73**	-0.13
Zn	**0.38**	**-0.42**	0.23	0.23	**-0.48**	**-0.57**	-0.15
Na	**0.39**	-0.11	0.12	**0.40**	**0.79**	-0.15	0.10
K	**0.42**	**0.44**	**-0.43**	0.17	-0.08	0.09	**-0.63**
Weighted average particle size	-0.24	**0.40**	**0.42**	**0.72**	-0.23	0.18	0.08

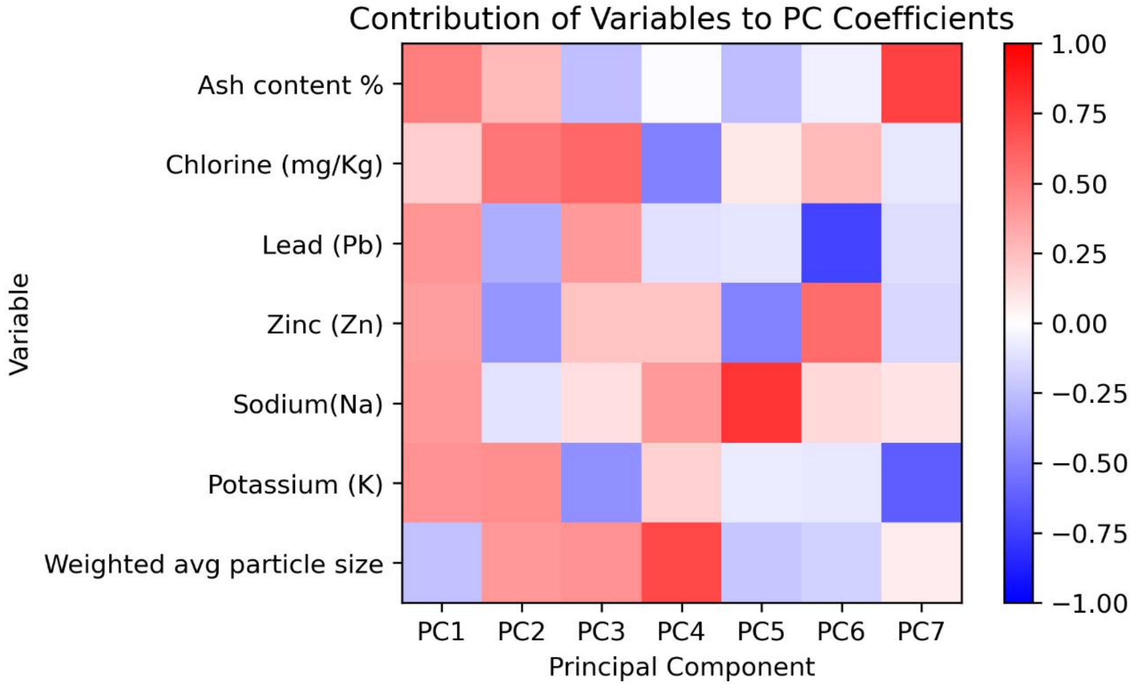

Figure 3: *Graphical representation of the coefficients for each principal component/variable pair.*

3.3 Random Forest Regression

Random forest regression was performed using all 30 data features (sample analysis measurement variables), to attempt to create predictive models for each of the six key components (Ash, Cl, Pb, Zn, K, Na). Error data for each of the six prediction cases is presented in Table 8. Scatter graphs showing predicted versus real values, using the testing data set, as well as the permutation importance for the prediction of each fuel component, are shown across Figure 4a-f and Figure 5a-f respectively.

Broadly speaking, whilst the model was able to find a good fit against the training data, it had a poor predictive accuracy when used on the testing data set. This was common across all the prediction targets, with only the prediction of ash having a moderate accuracy when using the testing data set with a MAPE of 19.05% as per Table 8. Across all other prediction targets, MAPE and MAE were around 3x worse with the testing data set versus the training data set. RMSE was similarly around 3x worse for the testing data versus the training data. To further contextualise how poor prediction accuracy was for most prediction targets, chlorine had a significant MAPE of 44.77% for the testing data set, indicating that the results were generally highly inaccurate. For the chlorine training data set however, MAPE was only 15.96%, which whilst not a small error, indicates that the model was able to create a reasonable "fit" during training. The coefficient of determination (R^2) values between training and testing data sets further shows the high degree of inaccuracy in the testing data cases. For ash prediction, R^2 was very high during training at 0.970, and moderate in the testing data case at 0.787. For all other prediction targets however, R^2 values whilst reasonably high during model training (typically around 0.9), were very low for the testing data sets, ranging from 0.173-0.545. This is again in line with what the other error measures indicated and implies a very poor correlation between the model prediction and real data when using the testing data set.

As a whole, the error data in Table 8 shows that the random forest algorithm was not able to find a generalised model to represent fuel chemistry that could be applied to "unseen" data, i.e. the testing data set. This is a different finding to the few other available studies that applied random forest to other biomass prediction challenges. For example, the work of Elmaz, et al. [18] displayed a 90% classification accuracy when predicting the overall fuel type (e.g. coal, wood, agricultural biomass). However, the prediction of ash and trace element contents is a fundamentally different problem to classifying fuel type, where there are well known differences between coals and biomass compositions such as volatiles or fixed carbon contents [2]. The lack of prediction accuracy here may be the result of the more heterogeneous data used here: instead of distinct data sets from different fuel types as in the work of Elmaz, et al. [18], a blended mixture of woods from various sources over a large timescale was analysed. More broadly, it is known that machine learning and artificial intelligence methods can achieve a higher degree of accuracy than that attained here when studying biomass fuels. For example, as highlighted in section 1.2, Li, et al. [20] predicted potassium content from flame spectroscopy data using a neural network approach and had an error of 6.34% with their best method.

The lack of accuracy or ability to construct a generalised model, is also evident across Figure 4a-f. The scatter plots of predicted values versus real values from the testing data set show the poor predictive power of the model across the cases. Again, only the prediction of ash content (Figure 4a) offers a moderately accurate prediction. If the trained model had a better understanding of the fuel fundamentals based on the data, then predictions with the testing data set would be better across the board. Therefore, this suggests that more data features (i.e. input measurement variables) are required to understand the data set.

Table 8: *Error data, for both the training and testing data sets, for each of the six model prediction targets.*

Prediction Target	**Error Evaluation Basis**	**MAE**	**MAPE (%)**	**MSE**	**RMSE**	**R^2**	**OOB Error**
Ash (wt.% dry)	Training Data	0.20	6.88	0.08	0.29	0.970	0.218
Ash (wt.% dry)	Testing Data	0.54	19.05	0.55	0.74	0.787	N/A
Chlorine (mg/kg dry)	Training Data	82.84	15.96	13521.75	116.28	0.895	0.774
Chlorine (mg/kg dry)	Testing Data	228.26	44.77	95198.09	308.54	0.245	N/A
Lead (mg/kg dry)	Training Data	7.71	14.67	135.77	11.65	0.916	0.3612
Lead (mg/kg dry)	Testing Data	20.14	39.17	888.60	29.81	0.437	N/A
Potassium (mg/kg dry)	Training Data	70.46	6.36	8707.18	93.31	0.936	0.471
Potassium (mg/kg dry)	Testing Data	195.19	17.24	63423.67	251.84	0.545	N/A
Sodium (mg/kg dry)	Training Data	16.54	10.54	685.27	26.18	0.911	0.631
Sodium (mg/kg dry)	Testing Data	50.52	30.57	8876.47	94.21	0.173	N/A
Zinc (mg/kg dry)	Training Data	8.68	13.23	238.96	15.46	0.925	0.540
Zinc (mg/kg dry)	Testing Data	27.12	38.23	2815.42	53.06	0.331	N/A

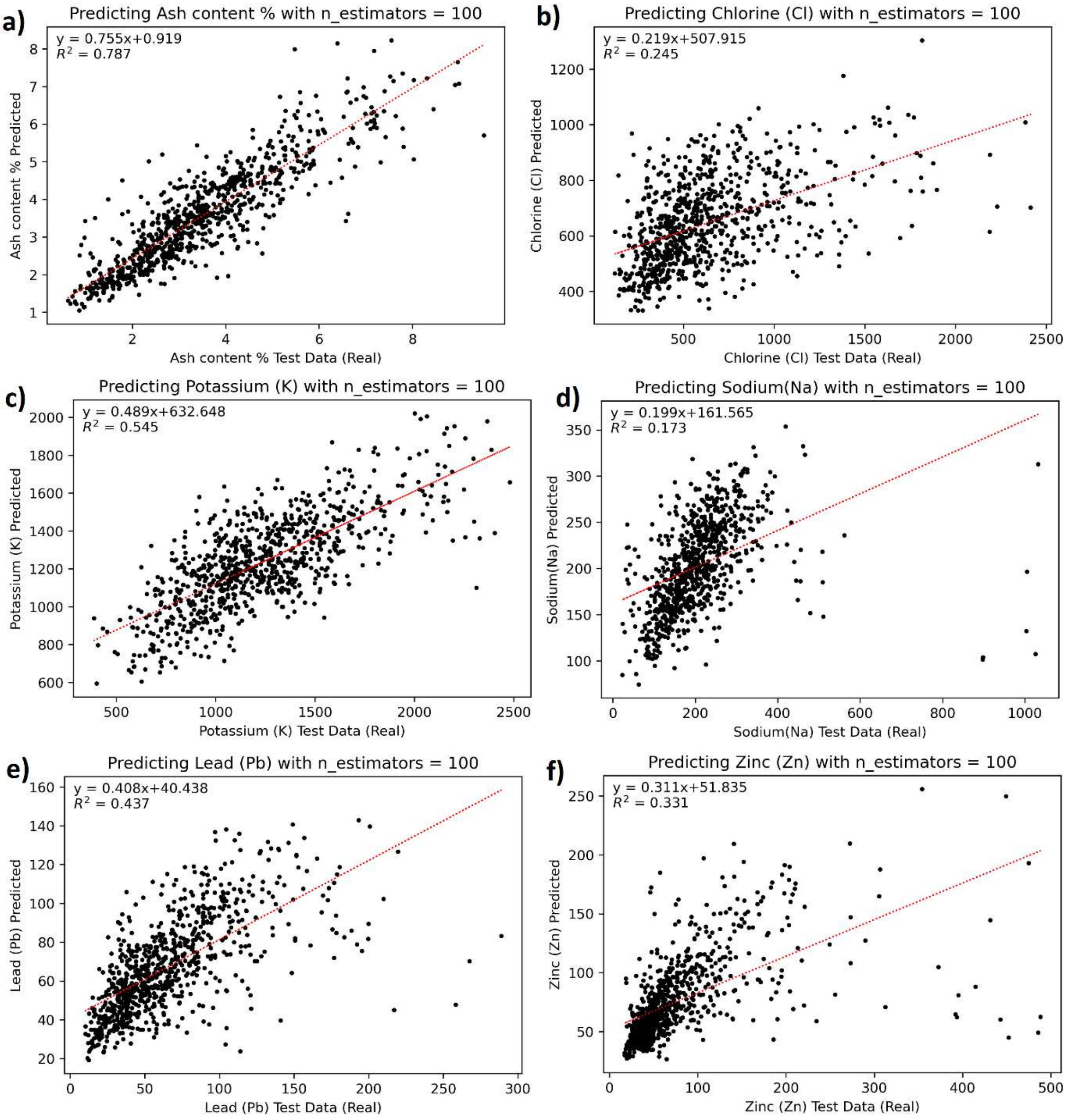

Figure 4: *Predicted values versus real data using the testing data set. a) Ash content. b) Chlorine. c) Potassium. d) Sodium. e) Lead. f) Zinc.*

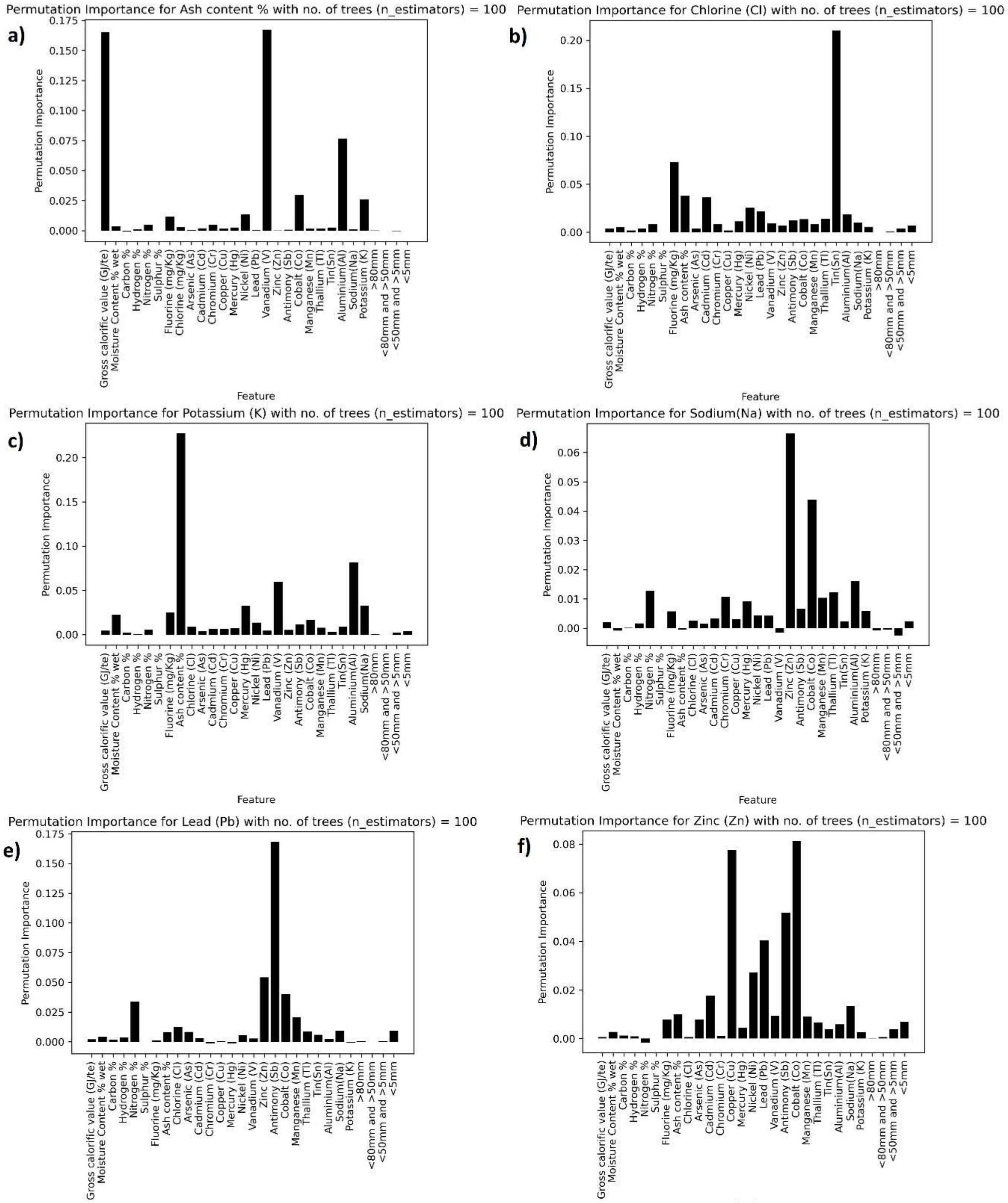

***Figure 5:** Permutation importance of input data features, when using the testing data set. a) Ash content. b) Chlorine. c) Potassium. d) Sodium. e) Lead. f) Zinc.*

From analysing the permutation importance charts (Figure 5), only a handful of components were identified as important to the model prediction of each prediction target, and these were only identified as moderately important (permutation importance ~0.15-0.25). As general prediction accuracy was weak, this is to be expected. It should be noted that permutation importance does not mean that a definitive correlation exists between two components, but that the component is important to the model and the accurate prediction of the target. It should also be considered that only ash showed a reasonable R^2 of 0.787 with the testing data set with the other prediction targets having R^2 <0.55. Therefore, for the other prediction targets, components with high permutation importance are likely far less important than they are for ash.

For the prediction of ash content (Figure 5a), GCV and vanadium had a moderate permutation importance of 0.16. For GCV, this would be expected, as GCV naturally has a degree of correlation with ash content (increased ash content means less potential for combustible components, hence lower GCV). In the case of vanadium, the underlying reason for this importance is less self-evident. Whilst it is derived mostly from inorganic sources, i.e. from waste wood ash as an extraneous contaminant [36], it is a very minor component with an average of 1.28mg/kg vanadium across 2786 samples of the fuel blend composition. Clearly, it was found to be a greater indicator of ash content than other components. This may be because out of all the trace elements present in the fuel, it is the most consistent indicator of overall ash content, as it is mostly derived from waste wood which typically has higher ash content and thus contributes more of the ash to the fuel blend [2].

For chlorine (Figure 5b), the only component to show a moderate permutation importance was tin with a value of 0.2. Tributyltin oxide is a commonly used wood preservative [37]. Therefore, tin could be a general indicator as to the presence of waste wood which would likely contain this preservative, with waste wood containing more chlorine than virgin wood (e.g. as plastic wastes and coatings) [2, 5].

For potassium (Figure 5c), ash content was seen to have a moderate permutation importance of around 0.25. This is broadly something that would be expected, as potassium is present within the ash content of all woods; being essential to wood growth [2]. It is notable that despite ash being important to the prediction of potassium, the reverse was not true (Figure 5a); i.e., potassium was not an important component for the prediction of ash content. This would suggest that for ash prediction, potassium had a far lower relative permutation importance than those components that were moderately important for ash prediction (GCV and vanadium).

For lead (Figure 5e), the only component with a moderate permutation importance was antimony with a value of 0.17. Antinomy is often blended in lead-based alloys. Therefore, this relationship is likely from the waste wood fraction of the fuel in combination with lead, with lead up to 10x more abundant in recycled wood than in virgin wood [36].

For sodium (Figure 5d) and zinc (Figure 5f), no component was identified as having even a moderate permutation importance, with the highest permutation importance seen in either case being ~0.08, versus upwards of 0.17 for the prediction of the other four key fuel components.

It is worth noting that prior studies, such as that of Vassilev, et al. [3] noted in section 1.2, which did not apply machine learning, have observed associations between certain elements in biomass ash. For example, Vassilev, et al. [3] identified woody biomass were commonly had higher levels of Ca, Mg, Mn and S versus other biomasses. Such associations were not identified by the random forest model here. The most likely reason for this is due to the model using large, heterogeneous data for a fuel blend, as opposed to small singular fuel data sets as were analysed and compared by Vassilev, et al. [3].

As mentioned previously, whilst the model was able to achieve a good fit with the training data for all prediction targets, it could not translate this accuracy to the unseen testing data set, with a large decline in accuracy in all cases. This suggests that the model was not sufficiently generalised and lacked sufficient data to identify the real trends present in the data. As previously mentioned, the work of Elmaz, et al. [18] shows that given sufficient fuel data, with data features that do have a strong correlation to the prediction target (proximate analysis data in the case of Elmaz, et al. [18]), random forest models can identify and classify fuels with good (~90%) prediction accuracy. Therefore, it stands to reason that with sufficient additional input information (i.e. data features) the model improve in predictive accuracy.

A further illustration of this is in the synthetic fuel exercise used to validate the model used in this study, discussed in detail in the Supplementary Data. In the first case, the model was used with a data set containing several random data features and two that have a 'y = mx' linear correlation with the prediction target. With no hyperparameter tuning, the model was able to predict ash content very accurately ($R^2 = 0.95$). The correlated data features were clearly evident in the permutation importance values. When the model was given entirely random data, the model achieved a "moderate" accuracy of $R^2 = 0.767$ during training. However, the model had very poor accuracy with near zero R^2 when applied to the unseen testing data set, which was also random. As was noted in this exercise, this showed that the model can fit to random data during training, but the decline in accuracy between training and testing is effectively a measure of how many correlating features are actually present in the data set, with the extreme case of no correlating features (all random data) resulting in very poor accuracy (near-zero R^2). Therefore, as can be seen from the results when predicting real fuel data across Table 8, Figure 4a-f and Figure 5a-f, the substantial decline in accuracy when moving from training to testing suggests that there are not a sufficient number of correlating data features present with the data set, hence an accurate and generalisable model cannot be constructed.

A further possible contributing factor, aside from a lack of correlating data features, is the heterogeneity of the fuel blend and data. It is a mixture of both virgin wood and recycled wood, with both sourced from several suppliers and mixed at point of use. This means that several different fuel sources are blended into one and then sampled/analysed, which may minimise or obscure correlations. In works where random forest has been successfully applied to fuels, such as that of Elmaz, et al. [18] as previously discussed, distinct fuel data sets were used as inputs rather than blends. Further to this, the fuel data was collected over long timescale (5 years). Previous studies such as that of Edo, et al. [5], have shown large changes in waste wood composition over such time scales. Variations in key component levels over time were also evident here in section 3.1, therefore this is also likely contributing to the difficulty in creating an accurate predictive model and discerning trends.

There are of course some general limitations to random forest regression that should be noted. One well known challenge is interpretability of how the model was created and how it functions, as without such understanding the model is essentially a “black box” [38]. This is a common critique and limitation applicable to most other alternative machine learning or artificial intelligence options, such as artificial neural networks. Here, the permutation importance of data features was evaluated in Figure 5 to better understand the important of each data feature to model creation and accuracy. Of course, interpretation of feature importance also requires a significant degree of domain expertise with regards to biomass fuels, to ensure that the relationships identified make sense in context. Other limitations, such as computational power requirements, were not a limiting factor. Rapid results were not required in this context, as the model was not to be implemented on a live process, but instead was to be used to retrospectively find relationships within the fuel data.

3.4 Recommendations

One of the largest sources of uncertainty in this study is the fuel data itself. As discussed, the recycled wood used by the power station is a highly heterogeneous fuel, frequently containing high levels of extraneous, non-wood contaminants, such as screws/nails, plastics, and other building/demolition waste which are not fully screened and removed prior to sampling and entry to the boiler. This makes fuel sampling challenging. An improvement to sampling, accounting for both the quantity and type of extraneous (non-wood) contaminants that are not currently analysed, would improve understanding of the overall fuel composition. The work of Edo, et al. [5] presents an example of hand sorting 1 tonne batches of waste wood to understanding the relative presence of different non-wood contaminants, such as stone or metals. Data from such an approach could be used in addition to the normal fuel analysis, to estimate the amount and composition of non-woody contaminants entering the boiler, that are not otherwise analysed during fuel analysis. This could be extended to also consider the breakdown of the types of waste wood in the stream (e.g. painted wood, coated particleboard), to then enable better selection of suppliers and sources.

As an additional point on the topic of fuel sampling, PAS 111 [39] is currently used in the UK as a specification for the sampling and grading of waste wood. Waste woods are classified from Grades A-D, where “A” is “clean” recycled wood through to “D” which is hazardous waste. However, waste wood grading is not uniform within Europe, with different countries using subtly different classifications [40, p. 43]. This is an area where a European standard may be of use. Moreover, the PAS 111 [39] grading system is largely qualitative, providing only approximate guidance to the types of wood and contaminants (e.g. particleboard, metals) that may arise. Improvements to fuel grading and improving sampling guidance such as by using the more detailed sampling guidance in BS EN 18135:2017 [41], would be a positive improvement for waste wood users.

With regards to the application of machine learning approaches such as random forest regression for biomass fuel quality prediction, it would be suggested that a focus is placed on narrower, clearly defined fuel quality data. The blended data set analysed here appeared to be too heterogeneous for the creation of an accurate predictive model. Creation of an accurate model would require the addition of more data features, such as the blend ratio, supplier, fuel source, breakdown of recycled wood types, etc. This would provide a better opportunity for the real underlying relationships to be identified. Identifying and recording all such important measurement data, which must be performed alongside the fuel analysis, is difficult to perform retrospectively. It is of course even more difficult to do so over such a long period (5 years), which limited the ability to further extend the model here. Moreover, increasing the amount of data recorded alongside fuel analysis will increase time and resource pressure on those collecting the data in the first instance. Once an accurate machine learning model is created however, mathematical equations to represent the underlying relationships can also be created. As strong relationships did not arise in the current study, the formation of mathematical equations was not attempted, as there would be little subsequent use for weakly correlated equations. If such equations can be formed in future works, it would be of significant value to researchers and industrial operators alike in using and understanding their fuels and blends.

4. Conclusion

From the analysis of the five-year fuel composition data set, there have been many quarters where key trace elements and ash content have been above the expected benchmark literature average levels. Ash content has exceeded literature average values for the blend in almost every quarter, whilst chlorine has exceeded average literature values for more than half of the quarters and would lead to exacerbated corrosion issues. This is likely due to the waste wood types and sources. Lead content was also seen to exceed literature average levels in almost every quarter to varying degrees, whilst older data showed high levels of zinc that have reduced over time. Likewise, ash levels have reduced during certain periods, indicating that ash and trace element levels can be better controlled.

Principal component analysis and random forest regression did not reveal any significant underlying relationships that could be used for composition prediction. Some minor, though expected, relationships were noted such as the decrease of GCV with increased ash content. The most likely reason for the inability to identify significant new relationships is the high heterogeneity of the fuel data, arising from the multiple fuel sources and suppliers used over a long (5 year) period at different plant operating points. A further contributing issue is the lack of other key data features within the data set, e.g. blend ratio. Applying these analytical techniques to narrower fuel data sets, e.g. only recycled wood from a specific supplier over a short time period, may give a better insight.

It is also recommended that fuel sampling and accounting for extraneous contaminants is improved. This is something that the industry in general may look to improve, as current UK guidance, e.g. PAS 111, does show areas for improvement.

5. Acknowledgements

J.D. Morris acknowledges funding support through EPSRC Centre for Doctoral Training in CCS and Cleaner Fossil Energy, grant no. EP/L016362/1.

6. References

[1] E. Johnson, “Goodbye to carbon neutral: Getting biomass footprints right,” *Environmental Impact Assessment Review,* vol. 29, no. 3, pp. 165-168, 2009.

[2] S. Vassilev, D. Baxter, L. Andersen and C. Vassileva, “An overview of the chemical composition of biomass,” *Fuel,* vol. 89, no. 5, pp. 913-933, 2010.

[3] S. Vassilev, C. Vassileva, Y.-C. Song, W.-Y. Li and J. Feng, “Ash contents and ash-forming elements of biomass and their significance for solid biofuel combustion,” *Fuel,* vol. 208, pp. 377-409, 2017.

[4] B. Gudka, J. Jones, A. Lea-Langton, A. Williams and A. Saddawi, “A review of the mitigation of deposition and emission problems during biomass combustion through washing pre-treatment,” *Journal of the Energy Institute,* vol. 89, no. 2, pp. 159-171, 2016.

[5] M. Edo, E. Björn, P.-E. Persson and S. Jansson, “Assessment of chemical and material contamination in waste wood fuels – A case study ranging over nine years,” *Waste Management,* vol. 49, pp. 311-319, 2016.

[6] N. Saqib and M. Bäckström, “Trace element partitioning in ashes from boilers firing pure wood or mixtures of solid waste with respect to fuel composition, chlorine content and temperature,” *Waste Management,* vol. 34, no. 12, pp. 2505-2519, 2014.

[7] Y. Alipour, P. Viklund and P. Henderson, “The analysis of furnace wall deposits in a low-NOx waste wood-fired bubbling fluidised bed boiler,” *VGB PowerTech,* vol. 12, pp. 96-100, 2012.

[8] Y. Alipour, P. Henderson and P. Szakálos, “The effect of a nickel alloy coating on the corrosion of furnace wall tubes in a waste wood fired power plant: The effect of a nickel alloy coating on waterwalls,” *Materials and Corrosion,* vol. 65, no. 2, pp. 217-225, 2014.

[9] M. Huron, S. Oukala, J. Lardière, N. Giraud and C. Dupont, “An extensive characterization of various treated waste wood for assessment of suitability with combustion process,” *Fuel,* vol. 202, pp. 118-128, 2017.

[10] E. Cesprini, G. Resente, V. Causin, T. Urso, R. Cavalli and M. Zanetti, “Energy recovery of glued wood waste – A review,” *Fuel,* vol. 262, p. 116520, 2020.

[11] C. Bale, E. Bélisle, P. Chartrand, S. Decterov, G. Eriksson, A. Gheribi, K. Hack, I.-H. Jung, Y.-B. Kang, J. Melançon, A. Pelton, S. Petersen, C. Robelin, J. Sangster, P. Spencer and M.-A. Van Ende, “FactSage thermochemical software and databases, 2010–2016,” *Calphad,* vol. 54, pp. 35-53, 2016.

[12] D. Lindberg, R. Backman, P. Chartrand and M. Hupa, “Towards a comprehensive thermodynamic database for ash-forming elements in biomass and waste combustion — Current situation and future developments,” *Fuel Processing Technology,* vol. 105, pp. 129-141, 2013.

[13] L. Breiman, “Random Forests,” *Machine Learning,* vol. 45, pp. 5-32, 2001.

[14] G. James, D. Witten, T. Hastie and R. Tibshirani, An Introduction to Statistical Learning, Springer Texts in Statistics, 2013.

[15] T. Hastie, R. Tibshirani and J. Friednman, The Elements of Statistical Learning, Springer, 2008.

[16] Elsevier, “Scopus,” Elsevier, 2020. [Online]. Available: https://www.scopus.com. [Accessed 15 October 2020].

[17] H. Ge, X. Li, Y. Li, G. Lu and Y. Yan, “Biomass fuel identification using flame spectroscopy and tree model algorithms,” *Combustion Science and Technology,* pp. 1-18, 2019.

[18] F. Elmaz, B. Büyükçakır, Ö. Yücel and A. Mutlu, “Classification of solid fuels with machine learning,” *Fuel,* vol. 266, p. 117066, 2020.

[19] TNO, “Phyllis2 Database for Biomass and Waste,” ECN, 2021. [Online]. Available: https://phyllis.nl/. [Accessed 15 September 2021].

[20] X. Li, C. Han, G. Lu and Y. Yan, “Online dynamic prediction of potassium concentration in biomass fuels through flame spectroscopic analysis and recurrent neural network modelling,” *Fuel,* vol. 304, p. 121376, 2021.

[21] P. Sakiewicz, K. Piotrowski and S. Kalisz, “Neural network prediction of parameters of biomass ashes, reused within the circular economy frame,” *Renewable Energy,* vol. 162, pp. 743-753, 2020.

[22] I. Jolliffe, Principal Component Analysis, Second Edition, London: Springer, 2012.

[23] Mathworks, “MATLAB R2019b,” Mathworks, Natick, Massachusetts, 2019.

[24] Scikit-Learn, “4.2 Permutation feature importance,” 2020. [Online]. Available: https://scikit-learn.org/stable/modules/permutation_importance.html. [Accessed 21 October 2020].

[25] Scikit-Learn, “3.3. Metrics and scoring: quantifying the quality of predictions,” Scikit-Learn, 2020. [Online]. Available: https://scikit-learn.org/stable/modules/model_evaluation.html. [Accessed 21 October 2020].

[26] J. Dibdiakova, L. Wang and H. Li, “Characterization of Ashes from Pinus Sylvestris forest Biomass,” *Energy Procedia,* vol. 75, pp. 186-191, 2015.

[27] Y. Niu, H. Tan and S. Hui, “Ash-related issues during biomass combustion: Alkali-induced slagging, silicate melt-induced slagging (ash fusion), agglomeration, corrosion, ash utilization, and related countermeasures,” *Progress in Energy and Combustion Science,* vol. 52, pp. 1-61, 2016.

[28] S. Enestam, R. Backman, K. Mäkelä and M. Hupa, “Evaluation of the condensation behavior of lead and zinc in BFB combustion of recovered waste wood,” *Fuel Processing Technology,* vol. 105, pp. 161-169, 2013.

[29] B. Ouvrard, J. Abildtrup, G. Bostedt and A. Stenger, “Determinants of forest owners attitudes towards wood ash recycling in Sweden - Can the nutrient cycle be closed?,”

Ecological Economics, vol. 164, p. 106293, 2019.

[30] H. Nielsen, F. Frandsen, K. Dam-Johansen and L. Baxter, “The implications of chlorine-associated corrosion on the operation of biomass-fired boilers,” *Progress in Energy and Combustion Science,* vol. 26, no. 3, p. 2000, 283-298.

[31] C. Forsberg, M. Broström, R. Backman, E. Edvardsson, S. Badiei, M. Berg and H. Kassman, “Principle, calibration, and application of the in situ alkali chloride monitor,” *Review of Scientific Instruments,* vol. 80, no. 2, p. 023104, 2009.

[32] B. Leckner, “Fluidized Bed Combustion: Mixing and Pollutant Limitation,” *Progress in Energy and Combustion Science,* vol. 24, pp. 31-61, 1998.

[33] E. Zabetta, M. Hupa and K. Saviharju, “Reducing NOx Emissions Using Fuel Staging, Air Staging, and Selective Noncatalytic Reduction in Synergy,” *Industrial & Engineering Chemistry Research,* vol. 44, pp. 4552-4561, 2005.

[34] G. Stubenberger, R. Scharler and I. Obernberger, “Nitrogen Release Behaviour of Different Biomass Fuels Under Lab-Scale and Pilot-Scale Conditions,” in *15th European Biomasss Conference and Exhibition*, Berlin, 2007.

[35] European Parliament, *Directive 2010/75/EU of the European Parliament and of the Council on Industrial Emissions,* Brussels, Belgium: Eruopean Union, 2010.

[36] J. Yan, A. Karlsson, Z. Zou and U. Edlund, “Contamination of heavy metals and metalloids in biomass and waste fuels: Comparative characterisation and trend estimation,” *Science of The Total Environment,* vol. 700, p. 134382, 2020.

[37] G. Sawyer and M. Irle, “Development of colour indicator techniques to detect chemical contamination in wood waste for recycling,” The Waste & Resources Action Programme, 2005.

[38] M. Aria, C. Cuccurullo and A. Gnasso, “A comparison among interpretative proposals for Random Forests,” *Machine Learning with Applications,* vol. 6, p. 100094, 2021.

[39] British Standards Institution, “Specification for the requirements and test methods for processing waste wood - PAS 111: 2012.,” British Standards Institution, London, 2012.

[40] M. Junginger, C. Hennig, T. Mai-Moulin, M. Edo and I. Johansson, “Transboundary flows of woody biomass waste streams in Europe,” IEA Bioenergy, 2018.

[41] British Standards Institution, “BS EN ISO 18135:2017 Solid Biofuels - Sampling,” British Standards Institution, London, 2017.