Int J Coal Sci Technol (2019) 6(2):207–217
https://doi.org/10.1007/s40789-018-0217-2

Understanding the impact of coal blending decisions on the prediction of coke quality: a data mining approach

Lauren A. North[1] · Karen L. Blackmore[2] · Keith V. Nesbitt[2] · Kim Hockings[3] · Merrick R. Mahoney[1]

Received: 14 February 2018 / Revised: 27 July 2018 / Accepted: 29 August 2018 / Published online: 7 September 2018

Abstract The accurate prediction of coke quality is important for the selection and valuation of metallurgical coals. Whilst many prediction models exist, they tend to perform poorly for coals beyond which the model was developed. Further, these models general fail to directly account for physical interactions occurring between the blend components, through the assumption that the aggregate properties of the blend are suitably representative of the overall behavior of the blend. To study this assumption, a parameter termed the vitrinite distribution category was introduced to directly account for the distribution of one of these commonly aggregated parameters, the vitrinite reflectance. The introduction of this parameter in a regression model for coke quality prediction improved the model fit. The vitrinite distribution category was demonstrated to provide new information about coal blending decisions, and was found to be capable of providing insight into the behavior of different blending structures. Residual analysis was applied to explore the behavior of the coke quality prediction model, with the vitrinite distribution category found to explain more than just the presence or absence of coals within a blend. This work provides the foundation of future studies in examining coal blending decisions, with the proposed parameter having the potential to be applied as part of a coke quality prediction model to optimize coal blending decisions.

Keywords Coke · Quality · Prediction · Self-organizing maps · Vitrinite reflectance distribution

1 Introduction

1.1 Coal, coke and the prediction of their properties

Metallurgical coke, derived from the pyrolysis of selected coals, plays several critical roles in the ironmaking blast furnace (Babich and Senk 2013; Bertling 1999; Biswas 1981). As a structural support, and source of permeability for the layers of softening iron materials, the selection of an appropriate coke can significantly influence operational stability (Bertling 1999). Due to the inability to directly measure the behavior of the coke within the blast furnace, several proxies to coke behavior under certain blast furnace conditions have been developed. Many prediction models have been developed to estimate the strength of coke produced from blends of coals (Díez et al. 2002; North et al. 2018a, b). These models are used to value and select coals, and to produce the desired coke. However, due to the complex, heterogenous nature of coke, the ability to predict its properties from the parent coals is a difficult task. Presently, there is no singular model which allows accurate prediction of the coke properties derived from coals of any coal basin (North et al. 2018b).

Early prediction models (Ammosov et al. 1957; Schapiro and Gray 1964; Schapiro et al. 1961) relied on the

✉ Lauren A. North
Lauren.North@uon.edu.au

[1] Centre for Ironmaking Materials Research, School of Engineering, The Newcastle Institute for Energy and Resources, University of Newcastle, University Drive, Callaghan, NSW 2308, Australia

[2] School of Electrical Engineering and Computing, The University of Newcastle, Callaghan, NSW 2308, Australia

[3] Coal Technical Marketing, BHP, 480 Queen Street, Brisbane, QLD 4000, Australia

characterization of behavior over the entire vitrinite reflectance distribution. However, the application of these models is limited to the coals they were derived from, and in part due to the seemingly arbitrary assumptions regarding the fusibility of macerals. The models were not able to be readily applied, or provide a reasonable level accuracy in their proposed form, at least without significant modification (North et al. 2018b). More recent models tend to rely on bulk, aggregated coal measures such as the mean maximum vitrinite reflectance, and thermoplastic behavior. However, due to the physical and chemical interactions occurring between the blend components (Sakurovs 1997, 2000; Sakurovs et al. 1994), and the non-additive nature of some of these measures (Lin and Hong 1986), these properties may be insufficient to capture these chemical interactions in an appropriate way.

1.2 Vitrinite reflectance and coal blending decisions

Coal rank is a prevalent feature in both coke quality prediction models and coal blending decisions, particularly in the form of the mean maximum vitrinite reflectance ($R_{v,max}$; North et al. 2018b). This prevalence is likely due to the relationship between coal rank and other key coal properties such as fluidity (Coin and Broome 1997; Ryan et al. 1998). The breakdown of the measured vitrinite reflectance into cumulative bins of 0.1% reflectance, termed vitrinoid types or V-groups, was popularized by Schapiro and coworkers in the 1960's (Schapiro and Gray 1964; Schapiro et al. 1961). These cumulative bins can be arranged graphically into a histogram, with the bars showing the percentage contribution of each V-group in the overall distribution. The shape of this histogram represents the distribution of reflectance in the measured spectrum. In the context of coal blending, vitrinite reflectance distributions are used to provide an indication of the similarity of the coals within blends (Bukharkina et al. 2012; Pearson 1991; Stankevich et al. 1998). In industrial use, completely overlapping ranges of reflectance, represented by a single peak within the histogram, is preferred by some coke producers over a multimodal vitrinite distribution (Choudhury et al. 2005; Yao 2008).

Although the concept of V-groups is an important factor in industrial coal blend preparation, the direct inclusion of V-groups in models that predict coke quality is limited. Several models (Bukharkina et al. 2012; Dash et al. 2005, 2012; Kishore et al. 2011; McKenzie et al. 1998; Suresh et al. 2012) consider an aggregate of a portion of the vitrinite distribution as an input parameter, in effect creating a weighted $R_{v,max}$. Kumar et al. (2008) reported that the vitrinite reflectance in the range V_9–V_{13}, a commonly used aggregate value, is an improved predictor of the measured Coke Strength after Reaction (CSR) than the mean maximum vitrinite reflectance. The standard deviation or petrographic non-uniformity of the distribution is another attribute used within models, that relates to V-groups and blending decisions (Bulanov et al. 2009; Stankevich and Bazegskiy 2013; Stankevich et al. 2008; Stankevich and Zolotukhin 2015). However, none of these methods are adequate in capturing the modality of the distributions, shown in Fig. 1. In the case of a distinctly bimodal distribution, neither the $R_{v,max}$, subset sum, or standard deviation approaches are able to suitably

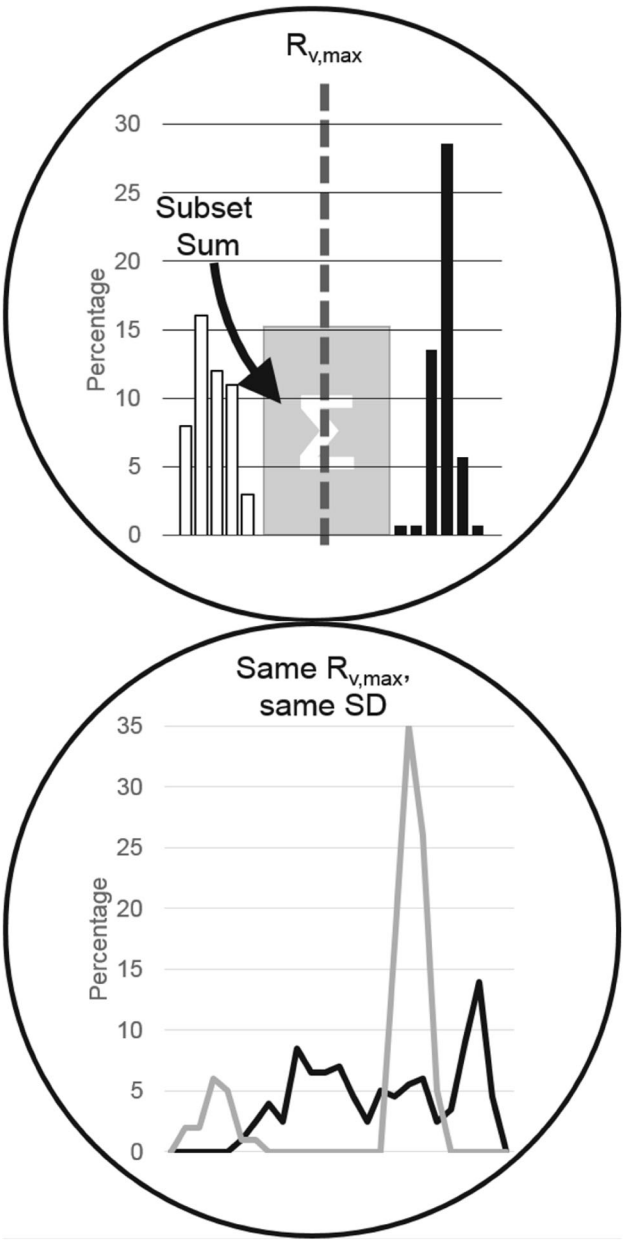

Fig. 1 Demonstration of the limitations of existing characterizations of vitrinite reflectance in determination of the modality of distributions, with both distributions shown having the same mean $R_{v,max}$ and SD

distinguish between a bi- or multi-modal distribution. These limitations of existing approaches suggest that alternative methods able to capture the diverse distributions of vitrinite reflectance may be useful in improving coke quality prediction models.

1.3 An introduction to knowledge discovery and data mining

This study makes use of two knowledge discovery techniques for the identification of patterns of coal blending, and for the exploration of model results, with the aim of improving coke quality prediction models. In general, knowledge discovery is the process of extracting new information, searching for patterns, and solving problems associated with large volumes of data using analytical tools (Han et al. 2011). In essence, the techniques used, typically termed machine learning, represent the region between traditional statistical analysis and artificial intelligence, although there is no clearly defined boundary between the techniques (Witten and Frank 2000). In terms of applications, data mining has been extensively utilized in the medicine, advertising, and manufacturing domains (Fayyad et al. 1996; Tsai 2012). Whilst not an exhaustive list of applications, the adoption within the metallurgical domain has been limited in comparison. This could be attributed to the domain requiring a certain level of knowledge in order to make meaningful contributions, or to the lack of general datasets suitable for data mining activities. In this regard, there is an opportunity to adapt these advanced analytical data mining techniques and further explore relationships within data sets that are not easily identified by linear methods.

This study seeks to develop an understanding of the influence of blending decisions on the prediction of coke quality. In particular, the influence of vitrinite reflectance distributions associated with blends is examined, through the development of a single parameter which captures the overall shape of the distribution. The impact of this parameter, termed the vitrinite distribution category (VDC), on coke quality prediction, is examined through regression modelling. An analysis of the residuals associated with regression modelling is applied to explore the information contained within the proposed VDC attribute.

2 Modelling framework

The following sections discuss the overall modelling approach, data sources, and calculations made as part of this study. The overall process is summarized in Fig. 2.

2.1 Background to vitrinite reflectance distribution analysis

As a single statistical measure, such as mean, does not adequately capture the shape of vitrinite distributions, a novel parameter that allowed robust classification of this behavior was developed (North et al. 2017). This parameter, the vitrinite distribution category (VDC), was derived using a data-mining algorithm known as a self-organizing map (SOM; Kohonen 1982). In essence, the VDC groups different coals, and blends of coals, into categories based on the full shape of their reflectance distribution. This method is shown schematically in Fig. 2 [for full details, see (North et al. 2017)] and in brief, comprises the following steps:

(1) Data cleaning—remove missing data, cumulative frequency not between 99 and 101, blends where a single coal represented more than 90% of the blend (i.e., only blends where used in this stage of the process)
(2) Offsetting—remove the influence of the $R_{v,max}$ by shifting each histogram left or right such that the $R_{v,max}$ was overlapped on the same location for each data point
(3) Conversion—convert the data to binary values by plotting as an image
(4) Determination of modelling parameters—iterative process to determine the parameters that need to be set for the SOM. See (North et al. 2017)
(5) Generating the classes—using the determined model parameters, a SOM was generated, with similar distributions grouped. For each grouping, a prototype vector, characterizing the group, is generated, with each data point assigned to a group based on the similarity to the prototype vector. This created the new categorical VDC variable, that captures a class of vitrinite reflectance distribution and can be used as a parameter in prediction models
(6) Fitting a regression equation—a regression model was generated for the case without using the assignment to VDCs (the *original* model), and a model including assignment to VDCs (the *V-group* model), improving the overall fit of the model. This analysis also included single coal data.

2.2 Application and modifications of previously described approach

The data used in this study is half V-group data, with bin sizes of 0.05% reflectance, taken from pilot coking oven experimental results, representing both Australian and some overseas coals, and their blends. In this study, the

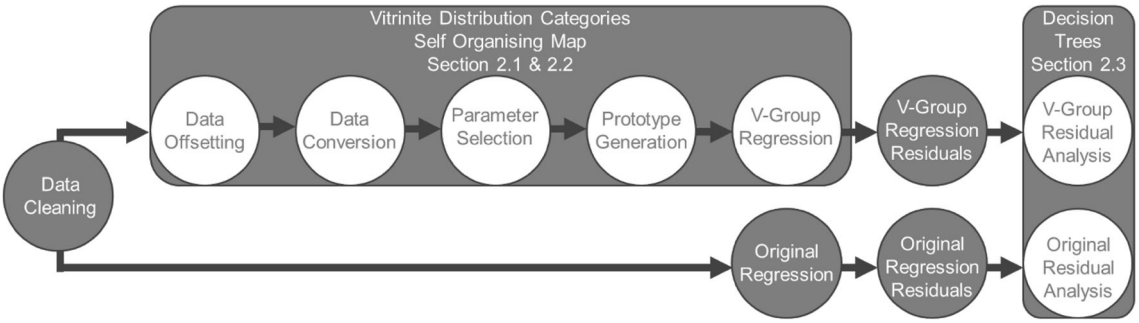

Fig. 2 Flowchart demonstrating overall modelling framework

method for measuring CSR is equivalent to the ISO standard method (International Organization for Standardization 2006), however a coke size of + 19–21 mm is used yielding no measurable difference in CSR results. To generate the categories contained within the VDC parameter, only blends were considered. As a modification to the approach described by North et al. (2017), additional new data was added to extend the training set. Data that had previously been filtered due to low measured CSR was also reintroduced. This change was adopted as it was considered that provided the V-group data itself was acceptable, then the CSR, and conditions under which the coke was produced would not affect the SOM itself. This extended the data set for the VDC parameter generation from 401 to 638 instances.

For the regression analysis, single coals, considered to be samples representative of a single coal brand or coal mine, were combined with the available coal blend data set. The data, however, was filtered prior to regression fitting to remove non-standard samples. Cases where the CSR was below 40 were also filtered, as in this region, measurement variability was considered to dominate. After the inclusion of single coals, and filtering accordingly, the total number of data points was 1039, of which 314 were blends.

The regression terms defined in the previous work (North et al. 2017) were also modified. The rank term that was previously used, $R_{v,max}$, was replaced with the blend volatile matter on a dry, ash free basis (VM_{daf}). The basis for this selection was the criticism that the average value of vitrinite reflectance has little physical meaning in the case of distinctly bimodal distributions. Conversely, the volatile matter term does physically represent the proportion of material that will be released upon heating, and can be verified by mass balance. Other attributes used in the regression analysis were other commonly used measures of coal quality, namely the modified basicity index (MBI), and the coal fluidity (logMF, the maximum fluidity expressed as a logarithm). The influence of the V-group distribution on the regression behavior was considered through analysis of both regression fit, as well as the residuals. Residual analysis was conducted using a decision tree, as discussed in the following section.

2.3 Residual analysis using decision trees

Interpretation of the behavior of the VDCs derived from the SOM analysis is important in further understanding the implications of blending decisions. In particular, a method for analyzing the residuals to better understand the combinations of coals that lead to under prediction is sought. Thus, in order to examine patterns of behavior associated with the blending of different coals, the residuals of both regression models were analyzed using a decision tree approach as discussed in the following section.

2.3.1 An introduction to decision trees, and the C4.5 algorithm

A decision tree is one class of data mining algorithm that allows classification of instances by their attributes (Tan et al. 2013). This particular approach is well suited to applications where the attribute to be classified is a categorical variable, such as the presence or absence of a coal within a blend. The algorithm generates a tree like flow chart structure, generating a series of splits in the data at branch nodes until a terminating leaf is reached, where the terminating leaf assigns the instance to one of the categorical variables (Han et al. 2011). The process of identifying these split points is dependent on the algorithm selected (Han et al. 2011).

In this analysis, the GNU General Public Licensed Weka 3.8.1 (Frank et al. 2016) machine learning software, and specifically the J48 decision tree algorithm, was used. The J48 decision tree is a Java based implementation of the C4.5 algorithm (Quinlan 1993), which uses the gain ratio measure to determine the value of a split (Tan et al. 2013). In brief, using the concept of entropy to represent the ordering of the data, this approach maximizes the difference between the resulting splits, which improves the

ordering of the data, reducing the entropy (Bramer 2016). Depending on the model set up, the tree may be evaluated during or after generation, with branches that add little value removed, in a process termed pruning. The C4.5 algorithm implements post pruning using a confidence measure of the classification error associated with each branch. If the decision node is unstable, then it is more likely to be removed during this process (Quinlan 1993).

2.3.2 *Implementation of decision trees*

In this work, the J48 algorithm in Weka was used. The key parameters, *confidenceFactor* and *minNumObj* have values of *U* and 2 respectively. The parameter *confidenceFactor* is used in the post pruning phase to determine the sensitivity of pruning, with the value *U* meaning that the tree is unpruned, whilst the *minNumObj* defines how many instances are required to generate a leaf node. These parameters were selected to grow the tree to maximum size.

Input and target data were generated from the dataset implemented in the regression. The input variables for the decision tree included the list of coals, encoded as a binary attribute, representing the presence or absence of that coal within the blend. The volatile matter was also implemented as a term to identify any systematic behavior associated with coal rank of blended components. The target variable was derived from each of the two regressions, also a binary attribute. This binary attribute was calculated from the difference between the residual value from the model, and the root mean square error calculated from the model. If the instance was underpredicted by more than the root mean square error of the model, it was classed as underpredicted. Conversely, if the instance was predicted within this range, it was classed as an acceptable prediction.

Comparison was made to the case where no rules are implemented and all instances are assigned to the dominant class (ZeroR in Weka), to assess the improvement of the model over this null hypothesis (Witten and Frank 2000).

It is noted that in this application, the decision trees generated are not intended to be used for a classification end goal; rather, they are used as an exploratory tool to aid in the assessment of the implications of the use of the VDC attribute in the prediction of coke quality.

3 Results and discussion

The following section firstly discusses the results of grouping the vitrinite reflectance distributions using the SOM algorithm, to produce the VDCs, and the implications on improving regression quality. The residuals of both regression models, with and without the grouped vitrinite reflectance distributions, are then analyzed using decision trees, and finally the associated links to vitrinite reflectance distribution are examined.

3.1 Implementation of the SOM

As described in Sect. 2.2, a self-organizing map was generated to group the vitrinite reflectance distributions. The SOM process generates VDCs, to which each instance is assigned. These VDCs are shown in Fig. 3. A tenfold cross validation was used to test the repeatability of the classification by evaluating the stability by which individual cases within the dataset are assigned to the same VDC during repeated runs of the process. The instances were assigned to the same category 87.6% of the time, indicating a high level of stability. Visual inspection of these VDCs notes a clear grouping of unimodal distributions in the top left corner of the map, corresponding to VDCs 4, 7, and 8. The remainder of the VDCs are associated with multimodal distributions, with varying proportions of high and low rank coal.

3.2 Regression analysis

3.2.1 *Original regression model*

As a benchmark for the remainder of the analysis, a regression which excluded the SOM classifications was completed using the method described in Sects. 2.1 and 2.2. The fit of this regression is displayed in Fig. 4. It is observed that a reasonable overall fit of the data is achieved, and that all inputs are statistically significant at the 5% level. However, it is noted that the fit of the data

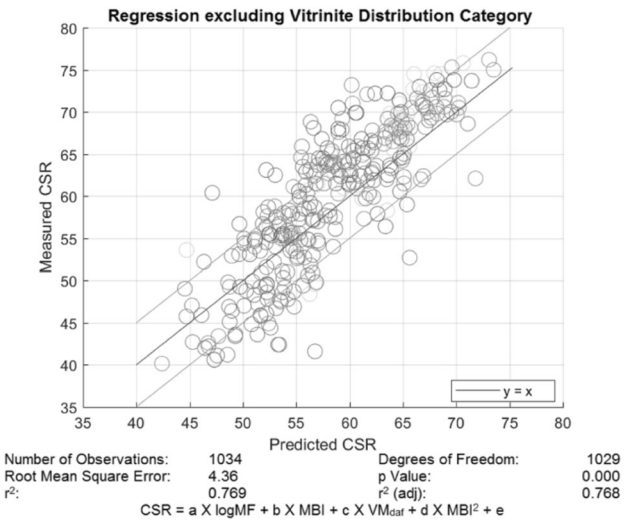

Fig. 3 VDCs derived from the self-organizing map. The numbers 1 through 9 are arbitrarily assigned to each grouping to provide a unique identifier

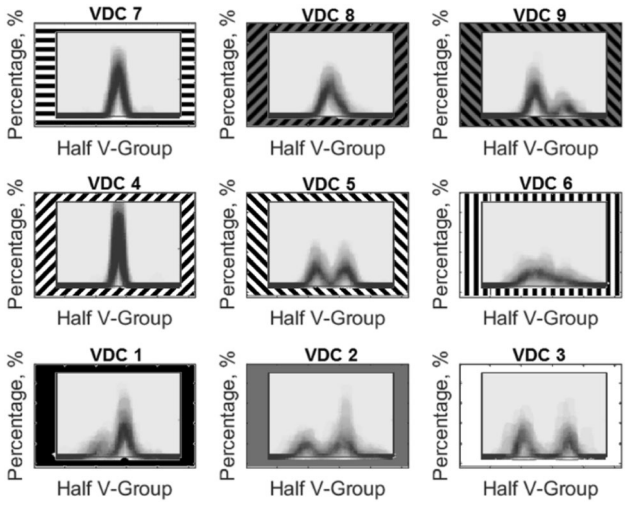

Fig. 4 Original regression blend fit and overall model statistics for both single coals and blends. Lines for $y = x$ and $y = x \pm 5$ are shown to aid comparison. For ease of interpretation, the figure displays only the results of blend fit, with the statistics reported developed on a model including both single coals and blends, excluding the VDC parameter

points representing blends are relatively poorly predicted when compared to the behavior of single coals.

Grouping the regression results by the associated VDC, as shown in Fig. 5, it is apparent that there is a systematic underprediction which is more prevalent at higher CSR values. This underprediction varies by each VDC. It is noted, however, that the fit of the model in the region of CSR below 50 is poor for both the single coals and blends. This poor fit is considered to be associated with the measurement variability that is known to increase as CRI increases (and hence CSR decreases; Menéndez et al. 1999). Despite the prevalence of Australian coking coals within the regression model, which could cause poor generalization to coals beyond this region, there is no apparent link between poor model fit and the presence of non-Australian coals.

3.2.2 *Regression including VDCs (V-group regression model)*

Each instance was grouped to one of the VDCs shown in Fig. 3. This grouping was assigned as a categorical variable within a regression model, using the method described in Sects. 2.1 and 2.2. The results of this V-group regression model are shown in Fig. 6.

Within this prediction, the coefficients (not shown) are similar between the original and the V-group regression models, and are statistically significant, with the exception of VDC 4, which is most similar in distribution to that of a single coal. It is observed that both the correlation coefficient and root mean square error improves with the introduction of the VDC attribute, and a more significant improvement in the prediction of blend behavior is also shown.

3.2.3 *Discussion of regression results*

Examination of the resulting coefficients associated with each of the VDCs shows a clear distinction between the behavior of single coals and blends. The coefficients associated with the VDCs are all positive, indicating that there is a systematic difference in behavior occurring due to the blending of coals and interaction between components that soften at different temperatures. Further, by sorting the coefficients associated with VDCs in ascending order, a trend associated with distribution types emerges, as shown in Fig. 7. It is evident that there are two distinct groups of behavior. Firstly, a blend containing many components generating a well overlapped distribution, behaves most like a single coal. As the distribution moves further away from well overlapped, towards distinctly multimodal distributions, the model coefficient, and hence difference in behavior from single coals, increases. Thus, the inclusion of a graphical form of vitrinite reflectance distributions in the regression model is capturing additional information about blending decisions. It is worth noting that inclusion of other statistical measures of the distribution, including the kurtosis and the standard deviation, are respectively not statistically significant or provide marginal improvement to the initial r^2 (0.769) and RMSE (4.36) values. Hence, this additional information that is captured by the graphical form of the vitrinite reflectance distributions derived via SOMs is unable to be captured by traditional statistical measures. The following section examines the nature of these blends with respect to the coal components they are made up of.

3.3 Regression residual analysis

Visual inspection of the results, particularly those that were underpredicted by the original model, suggests consistent underprediction associated with specific coals, despite the coal being well represented within the dataset. In order to explore the role that the VDCs are playing within the regression model, and whether any systematic behavior was present, the residuals from both the original model and the V-group regression model were analyzed. These residuals were analyzed according to the decision tree method described in Sect. 2.3.2.

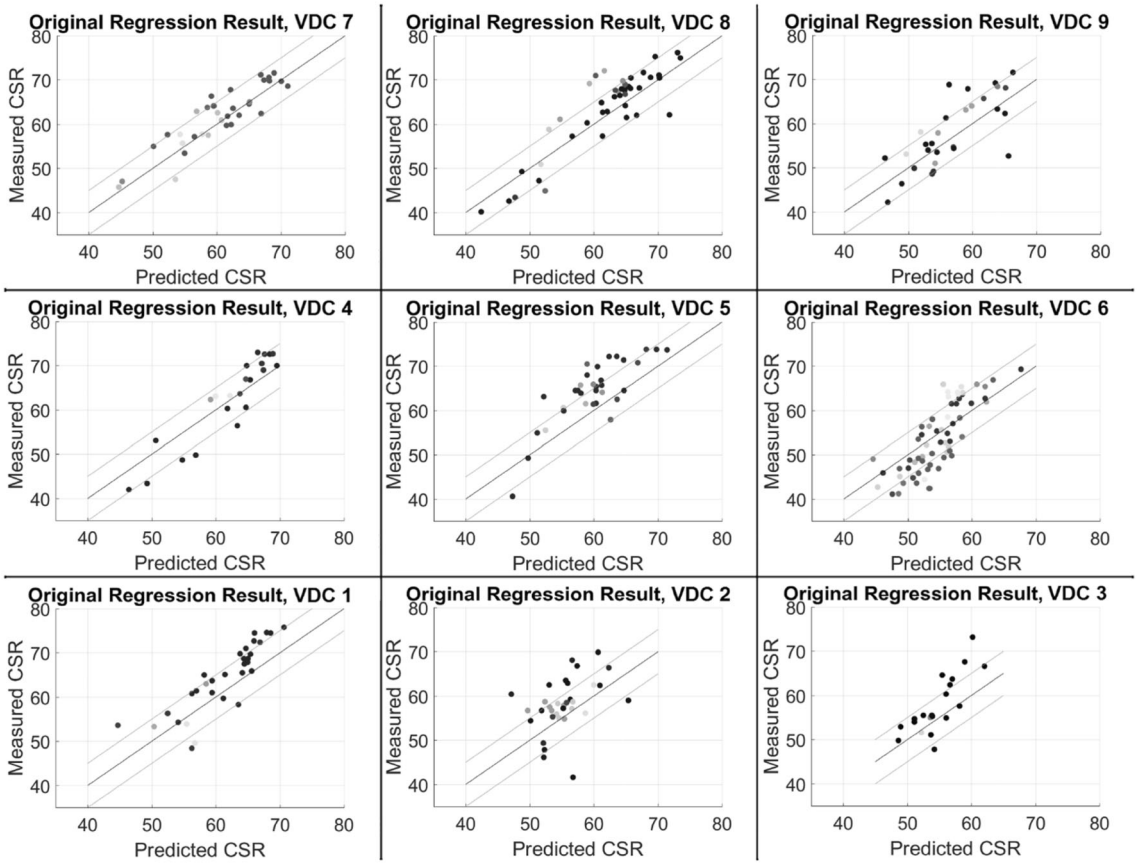

Fig. 5 Original regression fits, grouped by VDCs, and shaded by percentage Australian coals, where the darker the data point, the higher the proportion of Australian coal

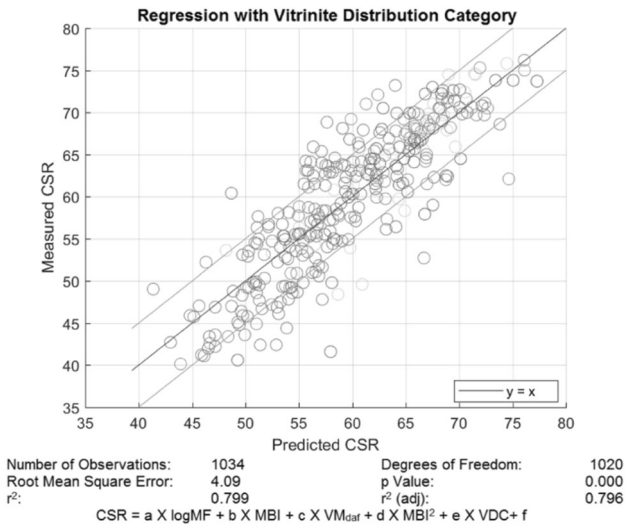

Fig. 6 V-group regression blend fit and overall model statistics for both single coals and blends. Lines for $y = x$ and $y = x \pm 5$ are shown to aid comparison. For ease of interpretation, the figure displays only the results of blend fit, with the statistics reported developed on a model including both single coals and blends, including the VDC parameter

3.3.1 Original regression residuals

In order to validate model behavior, residual plots were generated for each independent input variable, as well as for an alternate measure of thermoplastic behavior, the total dilatation. These residual plots, shown in Fig. 8, display random patterns, indicate no remaining systematic correlation with the key variables, and that the including the total dilatation within the model would add little value.

Figure 9 shows the decision tree that examines the behavior of underpredicted blends. As mentioned in Sect. 2.3, the intent of this decision tree is not as a predictive tool, but to explore whether patterns of behavior are associated with particular coals. This noted, the decision tree presented provides marginally better accuracy than the null hypothesis or uniform prediction to the majority class (ZeroR accuracy = 72.0%, full set accuracy 85.4%). It is evident from Fig. 9 that there are patterns associated with individual coals, rather than random classifications.

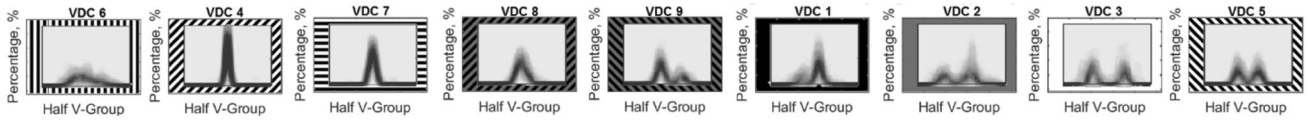

Fig. 7 VDCs sorted by model coefficient (smallest at left, greatest at right). The greater the model coefficient, the greater the underprediction by the original model. VDC numbering is consistent with that used Fig. 3

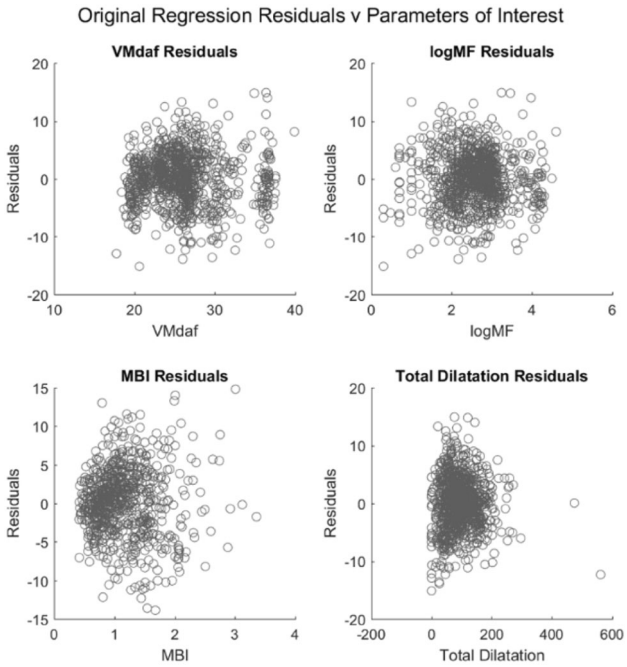

Fig. 8 Original model residual plots of independent input variables VMdaf, logMF, and MBI, as well as parameter of interest total dilatation

3.3.2 *Exploration of V-group regression residuals*

A similar analysis to the previous section was conducted on the V-group regression model residuals. Once again, these residuals were plotted against parameters of interest in Fig. 10, and analyzed according to the decision tree method described in Sect. 2.3.2, producing the resulting tree shown in Fig. 11. It is noted that as with the original model results, the V-group regression residuals show no patterns of behavior when plotted against the input parameters.

The decision tree presented provides minimally better accuracy than the null hypothesis or uniform prediction to the majority class (ZeroR accuracy = 84.1%, full set accuracy 89.5%). Visual inspection of the resulting decision tree suggests the presence of similar systematic behavior within the residuals to that identified for the original regression.

3.3.3 *Discussion of residual analysis*

As briefly discussed in the previous section, based on the residual plots in Figs. 8 and 10, the form of the regression

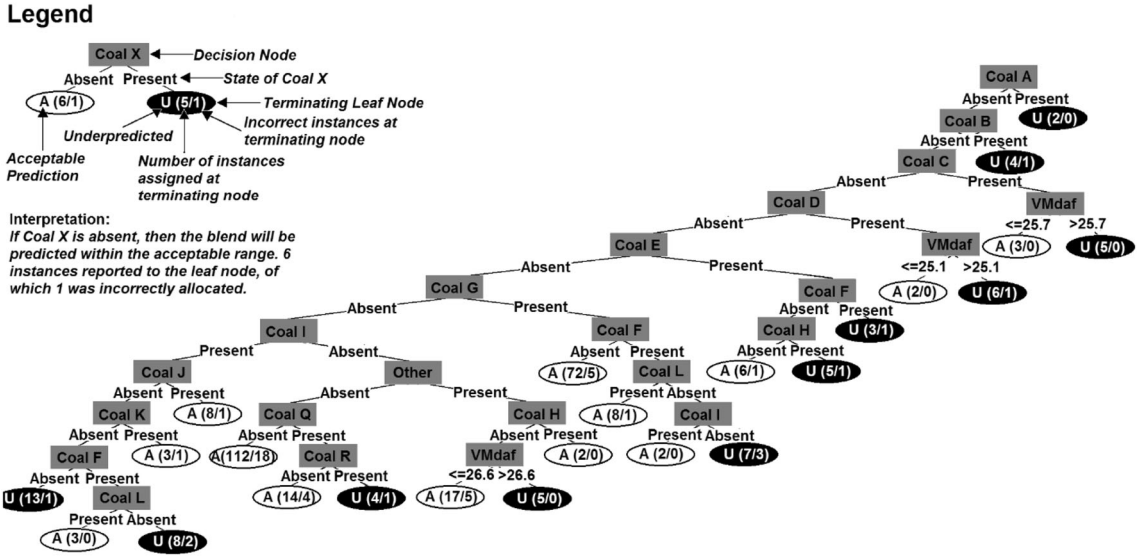

Fig. 9 J48 decision tree for underpredicted blends from the original model. Boxes indicate a decision point, whilst ovals are terminal nodes. A white node with black text represents a sample predicted by the original regression model within the RMSE, whilst a black node with white text represents a sample that is underpredicted by the original regression

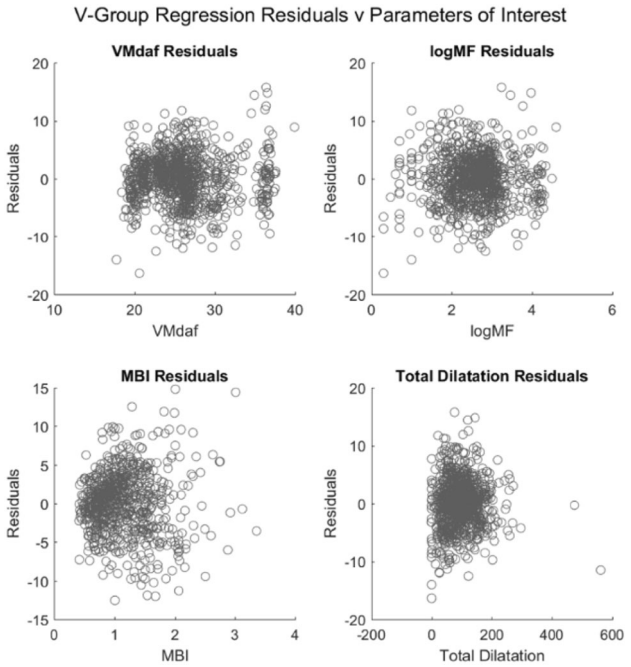

Fig. 10 V-group model residual plots of independent input variables VMdaf, logMF, and MBI, as well as parameter of interest total dilatation

of residuals in these residual plots remained approximately consistent with the original regression model, and the coefficients of the regression model changed minimally between the models. This suggests that the introduction of the VDC attribute is including new information within the regression model. Based on the ordering of coefficients in Fig. 7, it is suggested that blends where coals of dissimilar rank are used, will behave least similar to a single coal, implying that these coals together combine better than a single coal of comparable properties would.

Due to the overrepresentation of particular coals within the blends, and the counterintuitive pattern of underprediction identified in Fig. 7, it was initially believed that the VDCs may be heavily influenced by presence or absence of these coals, and hence that the new information that was being captured by these VDCs was coal related rather than related to the blending decisions as intended.

The residual analysis for both the original and V-group regression models, shown in Figs. 9 and 11 shows structure within the residuals. This structure takes the form of the presence or absence of particular coals that are consistently underpredicted. As a similar structure was found between both sets of residuals, suggesting that this behavior was not

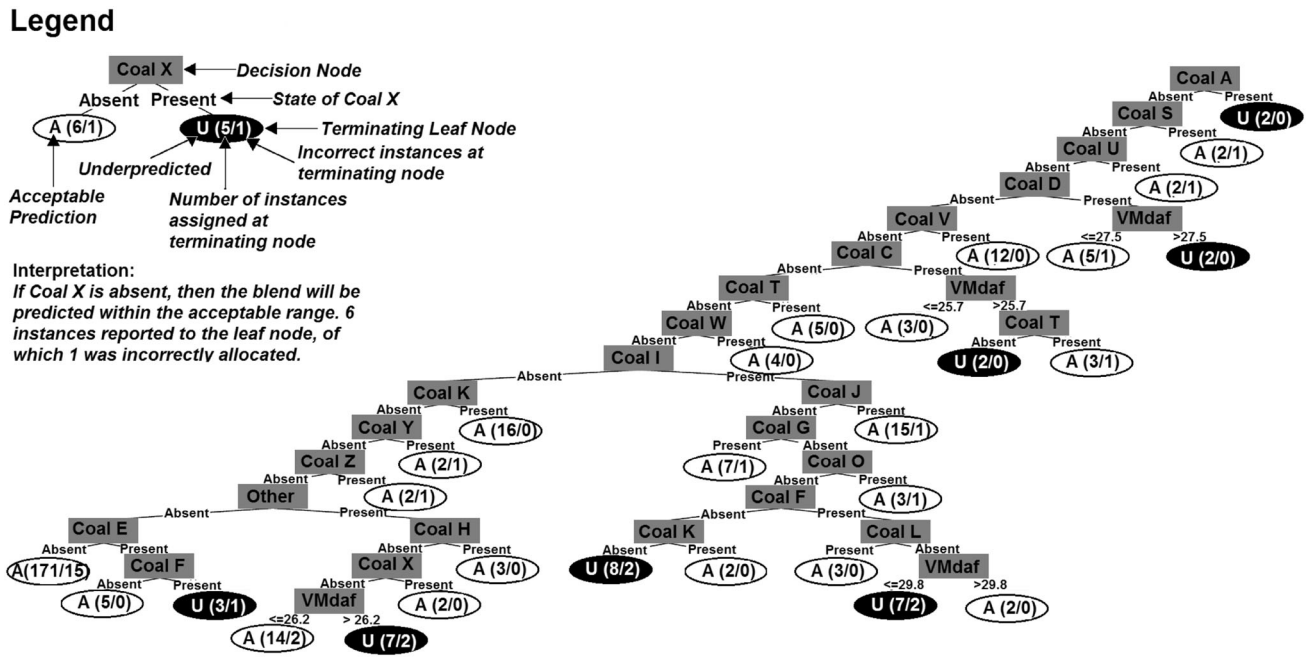

Fig. 11 J48 decision tree for underpredicted blends from the V-group model. Boxes indicate a decision point, whilst ovals are terminal nodes. A white node with black text represents a sample predicted by the original regression model within the RMSE, whilst a black node with white text represents a sample that is underpredicted by the original regression

model fitted to the data set is an appropriate fit. This form of the model is consistent with the main factors influencing coke strength after reaction identified by Díez et al. (2002). When the VDCs were added to the regression, the behavior captured within either model. Further, given the structural similarity of these residuals, it can be concluded that the VDCs are not representing the presence or absence of

particular coals within the blend, and are capturing new information.

As structure within the residuals was identified, there is an indication that there is some behavior associated with the properties of some coals that is not well captured by the regression model. As an underpredicted blend result implies that the blend is forming a better coke than the aggregate value of its components would suggest, both regression models are failing to capture the positive effect associated with the interaction of the blend components. Further study is required to identify the properties of the coals associated with underprediction.

4 Conclusion

This study has demonstrated the limitations of existing approaches of implementing the vitrinite reflectance distribution within a coke quality prediction model. The approach used within this work captured these distributions using a self-organizing map, generating a new parameter, the vitrinite distribution category (VDC), and identified that these distributions provide insight into blending decisions. The introduction of this parameter improved the fit of a regression model for coke quality prediction, and identified that blends of distinctly different rank coals were poorly fitted by models assuming the aggregate value of vitrinite reflectance. Residual analysis suggested that these blending behaviors are influenced by the underlying coals, and that none of the parameters considered within this modelling account for this behavior. Whilst the development of the VDC parameter improves the understanding of blending decisions on coke quality prediction, further study is required to identify and interpret the properties associated with particular coals that are affecting accurate prediction of coke quality.

Acknowledgements The authors gratefully acknowledge the constructive advice from Richard Roest, Hannah Lomas, and Kim Colyvas (The University of Newcastle).

Author Contributions The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

Funding We gratefully acknowledge the financial support of the Australian Coal Association Research Program (ACARP—Project C25077).

Compliance with ethical standards

Conflict of interest All authors declare that they have no conflict of interest.

References

Ammosov I, Eremin I, Sukhenko S, Oshurkova L (1957) Calculation of coking charges on basis of petrographic characteristics of coals. Koks Khimiya 2:9–12 **(in Russian)**

Babich A, Senk D (2013) Chapter 12: coal use in iron and steel metallurgy. In: Osborne D (ed) The coal handbook: towards cleaner production: volume 2: coal utilisation. Elsevier, Amsterdam

Bertling H (1999) Coal and coke for blast furnaces. ISIJ Int 39:617–624. https://doi.org/10.2355/isijinternational.39.617

Biswas AK (1981) Principles of blast furnace ironmaking: theory and practice. Cootha Publishing, Brisbane

Bramer M (2016) Principles of data mining. Springer, London

Bukharkina TV, Luk'yanov VL, Sal'nikova OY, Sokolovskaya EE (2012) Optimizing the batch composition at OAO moskoks. Coke Chem 55:138–141. https://doi.org/10.3103/s1068364x12040011

Bulanov EA, Krutenkov VG, Shilyakov AV (2009) Predicting the postreactive strength (CSR) and reactivity (CRI) of dry-slaked metallurgical coke on the basis of Audibert-Arnu dilatometry. Coke Chem 52:404–407. https://doi.org/10.3103/s1068364x09090063

Choudhury N, Boral P, Hazra S (2005) Coal petrography for prediction of coke quality of blended coals. In: International seminar on coal science and technology. Allied Publishers, p 224

Coin C, Broome A (1997) Coke quality prediction from pilot scale ovens and plant data. In: International coal conference, pp 325–333

Dash PS, Guha M, Chakraborty D, Krishnan S, Banerjee P (2005) Application of various techniques for predicting coke CSR from coal blend properties through laboratory scale experiments tata. Search 1:89–100

Dash PS, Guha M, Chakraborty D, Banerjee PK (2012) Prediction of coke CSR from coal blend characteristics using various techniques: a comparative evaluation. Int J Coal Prep Util 32:169–192. https://doi.org/10.1080/19392699.2011.640301

Díez MA, Alvarez R, Barriocanal C (2002) Coal for metallurgical coke production: predictions of coke quality and future requirements for cokemaking. Int J Coal Geol 50:389–412. https://doi.org/10.1016/S0166-5162(02)00123-4

Fayyad UM, Piatetsky-Shapiro G, Smyth P (1996) From data mining to knowledge discovery: an overview. In: Usama MF, Gregory P-S, Padhraic S, Ramasamy U (eds) Advances in knowledge discovery and data mining. American Association for Artificial Intelligence, Menlo Park, pp 1–34

Frank E, Hall MA, Witten IH (2016) The WEKA workbench vol fourth edition, version 3.8.1 edn. Morgan Kaufmann, Online appendix for "data mining: practical machine learning tools and techniques"

Han J, Pei J, Kamber M (2011) Data mining: concepts and techniques. Elsevier, Waltham

International Organization for Standardization (2006) ISO 18894:2006 coke—determination of coke reactivity index (CRI) and coke strength after reaction (CSR). International Organization for Standardization, Geneva

Kishore GS, Jagannadham G, Alma S (2011) Coal blend modeling and coke quality prediction studies—GIKIL's success story. In:

AISTech—iron and steel technology conference proceedings, pp 207–216
Kohonen T (1982) Self-organized formation of topologically correct feature maps. Biol Cybern 43:59–69. https://doi.org/10.1007/bf00337288
Kumar PP, Barman S, Ranjan M, Ghosh S, Raju VVS (2008) Maximisation of non-coking coals in coke production from non-recovery coke ovens. Ironmak Steelmak 35:33–37. https://doi.org/10.1179/174328107X174762
Lin M-F, Hong M-T (1986) The effect of coal blend fluidity on the properties of coke. Fuel 65:307–311. https://doi.org/10.1016/0016-2361(86)90288-7
McKenzie F, Hockings K, Livingstone P (1998) Modelling of coal blending to predict coke quality. In: Australian coal science conference. Australian Institute of Energy, Sydney, pp 101–106
Menéndez JA, Álvarez R, Pis JJ (1999) Determination of metallurgical coke reactivity at INCAR: NSC and ECE-INCAR reactivity tests. Ironmak Steelmak 26:117–121. https://doi.org/10.1179/030192399676997
North L, Blackmore K, Nesbitt K, Hockings K, Mahoney M (2017) A novel approach to coke strength prediction using self organizing maps. Paper presented at the DMIN'17–The 13th international conference on data mining, Las Vegas
North L, Blackmore K, Nesbitt K, Mahoney M (2018a) Methods of coke quality prediction: a review. Fuel 219:426–445. https://doi.org/10.1016/j.fuel.2018.01.090
North L, Blackmore K, Nesbitt K, Mahoney M (2018b) Models of coke quality prediction and the relationships to input variables: a review. Fuel 219:446–466. https://doi.org/10.1016/j.fuel.2018.01.062
Pearson D (1991) Probability analysis of blended coking coal. Int J Coal Geol 19:109–119
Quinlan J (1993) C4.5: programs for machine learning. Morgan Kaufmann, San Mateo
Ryan B, Grandsden JF, Price J (1998) Fluidity of western Canadian coals and its relationship to other coal and coke properties vol paper 1998-1. Geological Fieldwork 1997
Sakurovs R (1997) Direct evidence that the thermoplastic properties of blends are modified by interactions between the component coals. Fuel 76:615–621. https://doi.org/10.1016/S0016-2361(97)00049-5
Sakurovs R (2000) Some factors controlling the thermoplastic behaviour of coals. Fuel 79:379–389. https://doi.org/10.1016/S0016-2361(99)00173-8
Sakurovs R, Lynch LJ, Maher TP (1994) The prediction of the fusibility of coal blends. Fuel Process Technol 37:255–269. https://doi.org/10.1016/0378-3820(94)90019-1
Schapiro N, Gray RJ (1964) The use of coal petrography in coke making. J Inst Fuel 11:234–242
Schapiro N, Gray RJ, Eusner G (1961) Recent developments in coal petrography blast furnace, coke oven and raw materials committee. Proceedings 20:89–112
Stankevich AS, Bazegskiy AE (2013) Optimizing coke production at OAO EVRAZ ZSMK on the basis of the available coal. Coke Chem 56:364–371. https://doi.org/10.3103/s1068364x13100098
Stankevich AS, Zolotukhin YA (2015) Determining the technological value of coal on the basis of coke-quality predictions. Coke Chem 58:233–244. https://doi.org/10.3103/s1068364x1507008x
Stankevich A, Chegodaeva N, Vens V, Cheremiskina A (1998) Optimization of the coal-blend composition and prediction of coke quality based on chemico-petrographic parameters. Coke Chem 1:14–24
Stankevich AS, Gilyazetdinov RR, Popova NK, Koshkarov DA (2008) Predicting CSR and CRI of coke on the basis of the chemical and petrographic parameters of the coal batch and the coking conditions. Coke Chem 51:357–363. https://doi.org/10.3103/s1068364x08090056
Suresh A, Ray T, Dash PS, Banerjee PK (2012) Prediction of coke quality using adaptive neurofuzzy inference system. Ironmak Steelmak 39:363–369. https://doi.org/10.1179/1743281211Y.0000000087
Tan PN, Steinbach M, Kumar V (2013) Introduction to data mining: pearson new international edition. Pearson Education Limited, London
Tsai H-H (2012) Global data mining: an empirical study of current trends, future forecasts and technology diffusions. Expert Syst Appl 39:8172–8181. https://doi.org/10.1016/j.eswa.2012.01.150
Witten IH, Frank E (2000) Data mining: practical machine learning tools and techniques. Academic, San Diego
Yao B (2008) The function reflectance distribution map in guiding coal blend for cokemaking. Fuel Chem Process 39:11–19 **(in Chinese)**