Application of PCA-Kmeans method-based BP neural network to the prediction and optimization studies in S ZORB Sulfur Removal Technology

Xiaoyi Geng
University of Shanghai for Science and Technology https://orcid.org/0000-0002-9096-9717
Guangcheng Zhang
University of Shanghai for Science and Technology https://orcid.org/0000-0003-3460-098X
Xin Wang (wangxinshiyun@126.com)
University of Shanghai for Science and Technology https://orcid.org/0000-0001-9108-3478
Bingyan Song
University of Shanghai for Science and Technology https://orcid.org/0000-0002-0219-5919
Yu Chen
University of Shanghai for Science and Technology https://orcid.org/0000-0002-6334-595X

Research article

Keywords: PCA-Kmeans method, BP neural network, S ZORB SRT, Random walk iteration

Posted Date: December 28th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-132338/v1

Abstract

In this paper, the modeling of predicting the gasoline octane number and sulfur content in S ZORB Sulfur Removal Technology (SRT) is established. In the modelling, the principal component analysis (PCA) and unsupervised K-means clustering algorithm were initially integrated together to determine the key variables that affect the octane number and sulfur content of the product. With the selected key variables, the backpropagation neural network prediction models of the product octane number and sulfur content were established, trained and tested. Moreover, the mean accuracy of the prediction error within 0.15 and 0.3 were 94% and 99%, respectively. Besides the prediction of output of the S ZORB SRT Reactor, a multi-variable random walk optimization method was also proposed and investigated to reduce the octane loss, which was expected to be reduced by more than 30%, during desulfurization of fluid catalytic cracking gasoline in the S ZORB SRT Reactor, meanwhile the sulfur content stayed relatively stable which was less than 5 ppm. The results of the proposed models are reliable and could be applied into the real industrialization, which are beneficial with both the efficiency of economy and environmental protection.

Highlights

- The modeling of predicting the gasoline octane number and sulfur content in S ZORB Sulfur Removal Technology (SRT) is established.
- The principal component analysis (PCA) and unsupervised K-means clustering algorithm are innovatively integrated together to determine the key variables that affect the octane number and sulfur content of the product.
- The multi-variable random walk optimization method is proposed and investigated.
- The proposed model not only can reduce the product octane loss value but also keep the product sulfur content below 5 ppm.
- The results of the proposed model are reliable and could be applied into the real industrialization, and suitable to be extended to other complex application scenarios.

1. Introduction

At this stage, one-third of the world's commercial gasoline comes from catalytic cracking units, and the produced catalytic cracking gasoline contains more than 90% of sulfur [1, 2]. The exhaust emissions from the combustion of high-sulfur gasoline will seriously damage the atmospheric environment and human health. S ZORB Sulfur Removal Technology, a sorbent-based sulfur-removal method is widely used in gasoline desulfurization [3, 4]. Under suitable pressure, temperature, and hydrogen conditions, adsorbents such as zinc oxide and nickel oxide are used in fluidized bed reactors. The sulfur contained in raw gasoline is adsorbed in the form of metal sulfide to produce gasoline components with low sulfur content (10 ppm) [5–8]. S ZORB SRT Reactor can produce gasoline with low octane loss and low sulfur value, which is also an important standard for evaluating desulfurization process [9]. As of 2019, 32 domestic installations have been reported, and S ZORB SRT Reactors has been placed into production in China, which effectively guarantees the domestic supply of clean gasoline energy. However, using this technology for desulfurization, the octane number of the product will inevitably decrease. At present, the octane loss of most S ZORB SRT Reactors is 0.5–1.5 units, which results in economic losses and reduces the production of high-grade gasoline. Taking the Chinese gasoline market in October 2020 as an example, the research octane number (RON) value of 92# car gasoline and 95# differs by 3 units, and the price difference is 300 yuan/ton, which is equivalent to 1 unit of RON value at 100 yuan/ton. Therefore, reducing the octane loss during S ZORB SRT process is important for enterprises to increase profits and improve energy utilization [10].

Many models have been proposed to explain the principle of the process. In the classic chemical process modeling, the loss of the octane number could be reduced [9, 11–14]. Given the late proposal of S ZORB SRT technology, many operating procedures and complex related mechanisms remain to be studied, and establishing a model from the reaction mechanism is difficult. However, some scholars have made corresponding contributions. Bezverkhyy et al. used thermogravimetric analysis to explore the reaction kinetics of thiophene on the adsorbent under laboratory conditions and divided the reaction of thiophene on Ni/ZnO into three different stages: rapid adsorption stage, surface reaction control stage, and solid-phase diffusion stage. This method gives a principal analysis of the working process of the adsorbent and establishes a model [15]. Schmidt et al. studied the effect of NH_3 and HCL on the surface activity of the catalyst [16]. Qiu et al. used X-ray photoelectron spectroscopy and X-ray diffraction to study the performance conversion between the catalyst and regenerated catalyst [17]. Jia et al. proposed a reactor modeling method on the basis of the process mechanism, which divided FCC gasoline into five lumps and built a reaction kinetic model and an octane number correlation model on the corresponding basis [18]. However, the use of chemical process modeling focuses on the mechanism analysis of a certain stage of production or the content of a certain component and on some parts and tends to overlook the overall desulfurization.

In the recent years, data science has been paid more and more attention, not only because data association methods can comprehensively consider the impact of multiple variables, but also classification and regression algorithms such as neural networks, support vector machines,

random forests and etc. can establish correspondences between high-dimensional spaces and nonlinear multi-variables. It was stated that the above data method has played an important role in the chemical industry and can be used to reduce chemical mixture risk, predict potential drug-drug interactions, automatic chemical design and so on. [19–24]

With regard to gasoline components prediction, the research is based on the theoretical basis of spectral analysis. After obtaining the spectral data of gasoline, different data fitting algorithms are used to predict the octane content of gasoline [25–28]. However, no paper has proposed a method to obtain the gasoline components by establishing a model of S ZORB SRT process.

Given the lack of mathematical model for the desulfurization of FCC gasoline, previous optimization of gasoline desulfurization used the controlled variable method to study the variables that may affect the product index. For example, Xiong et al. studied the effects of space velocity, adsorbent bed temperature, adsorbent roasting temperature, and roasting time on the desulfurization rate of gasoline [29]. However, this method cannot achieve optimization at the same stage of multiple variables, ignoring the interconnection among variables.

Therefore, prediction models will be established for the sulfur content and octane number of the S ZORB SRT reaction and achieve optimization at the same time of multiple variables through the random walk iterative optimization algorithm. Using the data method to model the S ZORB SRT reaction, the data contains more than 300 variable information such as raw material properties, catalyst properties, and device process control variables. Directly use raw data for modeling will lead to complications of multicollinearity and other false correlations in model variables, which will affect generalization ability. Therefore, data dimensionality reduction is necessary to extract the key variables. The common methods include principal component analysis (PCA), clustering, and manifold learning [30–33]. In these papers, PCA is used to reduce the dimension of algebra, and K-means algorithm is used to cluster variables. Due to the sensitivity of the K-means cluster method to the initial positions, which determines the final cluster result. After performing PCA analysis, the result is then passed for K-means algorithm as initial positions of cluster centers.

In this paper, a PCA-Kmeans method-based BP neural network model of the S ZORB SRT reactor will be established. Based on the model, an optimization method for operating variables is proposed to reduce the product octane loss value by more than 30%, meanwhile keep the product sulfur content below 5 ppm. First, PCA and K-means clustering algorithm are used to select the representative key variables from many physical and operating variables during S ZORB SRT process. Then, the key variables are used as input to construct a neural network prediction model and test the performance of the model. Finally, based on the S ZORB product sulfur content and octane number prediction model, a multivariate optimization to reduce sulfur content and octane number loss is proposed. The algorithm can improve the octane content of desulfurized gasoline.

2. Modeling Analysis Of The S Zorb Srt

This article takes the S ZORB SRT reactor of a petrochemical company in China as an example and establishes a model through data samples collected by the reactor. The collected data are the continuous operation of the reactor from April 2017 to May 2020. Each data sample includes 354 operating variables and 7 raw material properties (sulfur content, octane number, saturated hydrocarbons, olefins, aromatics, bromine number, and density), product property variables (sulfur content and octane number), spent adsorbent variables, and regenerated adsorbent variables. Except for the product property variables, which are the model prediction results, other variables all affect the product of the reactor.

PCA was firstly to process the original independent variable matrix to output the principal components. Then, K-means clustering was performed on the principal components obtained in the previous step and the variables closest to the cluster centers in each cluster of the clustering results were used as the initial cluster centers and key variables, respectively. Moreover, the BP neural network models could be established to predict the sulfur content and octane content of the desulfurized gasoline in the S ZORB SRT reactor.

2.1 Extraction of principal components by PCA

PCA was first introduced by K. Pearson on non-random variables, and the amount of information reflected by variables with correlations has a certain overlap [34]. Using the linear algebra technique of continuous attributes, new attributes (principal components) in the data can be obtained [35, 36]. These attributes are linear combinations of the original attributes and are mutually orthogonal without overlapping information. Therefore, PCA is often used to reduce dimensionality in mathematics [37]. The key variables of the S ZORB SRT reactor can be divided into raw material property variables, adsorbent property variables, and operating variables. This classification bases that changing raw material properties and adsorbent properties during optimization is more complicated than operating variables. Therefore, this article has classified all variables into three categories: raw material properties, adsorbent properties, and operating variables. Data preprocessing is performed on the original data, including the following preprocessing processes: (a) eliminating the abnormal values based on the range of each operating variable; (b) the utilization of the Pauta criterion proposes outliers in the data; (c) replacing the missing values with the average of the 2 h data before and after the operation; (d) deleting the variables with many missing data. Preprocessing has removed eight operating variables with many data missing in the operating variables. Then, the PCA method is used to process the three types of variables.

The number of principal components is calculated when the cumulative contribution rate of the raw material properties and adsorbent properties is greater than 0.9, and the number of principal components is calculated when the cumulative contribution rate of operating variables is greater than 0.85. The variance contribution rate of each component is shown in Fig. 1, and the calculation results are shown in Table 1.

As shown in the table, the number of main components of raw material properties and adsorbent properties accounts for more than 75% of the variables, whereas the number of main components of the operating variables only accounts for 10% of the variables. This result is consistent with the actual situation. The nature of the raw materials has a small correlation, whereas the operating variables have a high correlation, such as the input pressure and output pressure of the container.

2.2 K-means clustering

K-means clustering algorithm is a clustering algorithm based on partition, which divides the data into valuable groups (clusters) according to the relationship and information of the objects described by the data [38–40]. The goal of division is as follows: objects in the groups are similar (related), and objects in different groups are dissimilar (irrelevant). Therefore, K-means clustering algorithm is often used in data processing, such as feature extraction and sample classification [41]. The key of the K clustering algorithm is to determine the K value. The principal component of the PCA result is used as the initial clustering center of the K clustering algorithm [42]. The theoretical basis for this approach is that the correlation between the clusters with the principal component as the initial clustering center is weak, and the correlation between the variables within the cluster is strong. The original variable matrix information can be represented by the clustering results.

In the clustering process, the initial cluster centers are 30 principal component variables obtained by PCA, and the cluster proximity measure is set to Euclidean distance, and the sum of squares of errors (SSE) is used as the objective function to determine clustering. The function formula is shown in Eq. (1).

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist(c_i, x)^2 \qquad (1)$$

The K-means clustering algorithm is used to cluster variables, and the partial results of cluster are shown in Fig. 2, Table 2. At present, PCA integrated with K-means was used to obtain the key variables that are representative and contain most of the information in the original variable matrix. The key variables in the properties of raw materials are hydrocarbon content, octane content and so on. The spent catalysts are the key variables in the catalyst properties. This finding may indicate that the spent catalysts are directly related to the process of adsorbing sulfur compounds at this stage. The key variables in the operating variables are space velocity, temperature, pressure, and hydrogen concentration, which conform to the S ZORB SRT. In addition, we will analyze the correlation of the 30 key variables. The results are shown in Fig. 3. If the colors of the selected key variables are lighter, then the correlation is weak, and the differences between clusters are evident, indicating that the clustering effect is remarkable.

2.3 BP neural network model

The above-mentioned algorithm is used to obtain the key variables and establish the prediction model. The Pearson correlation analysis result of the sulfur content and octane number of the product is 0.208 [43], therefore, these two properties are weakly correlated. These two properties can be used as two types of variables [44–46].

The network initialization training parameters are defined as follows: the number of neurons in the input layer is 30, and the number of hidden layers of the prediction model is determined to be 18 through empirical formulas and multiple tests, and a neural network with a topology of 30-18-1 is obtained (as showed in Fig. 4). The transfer function of the hidden layer is logistic, and the transfer function of the output layer is purelin. The initial learning rate of the neural network is 0.001, and the learning rate change mode is adaptive. The minimum convergence error of the training target is 0.001, and the maximum number of training times is 1000. The data was divided into 80% training set and 20% test set.

It is defined as the correct prediction criterion that the differences between the predicted result and actual value is less than 0.15. The ratio of the predicted correct variables in the test set to the total number of test sets is the prediction accuracy rate. After many parameter adjustments, the final prediction accuracy rate of RON and Sulfur content is shown in Figs. 5 and 6, and the performance of the neural network is shown in Table 3.

As shown in Figs. 5 and 6, Table 3, the model prediction results almost coincide with the actual results. The average accuracy rate reaches 94% when the error value does not exceed 0.15 and reaches 99% when the error value does not exceed 0.3. It is accurate and can be used in the next optimization. In addition, the octane number fluctuates by about 4 units, whereas the sulfur content fluctuates slightly, mostly stable at approximately 3.2 ppm. This result indicates that the octane number is easily affected by the input variables, and the sulfur content is robust, which provides the basis for the following optimization method.

3. Model-based Optimization Method Of Key Variables

In actual production, we hope to reduce the loss of octane number and sulfur content to increase economic benefits. Hence, we propose the optimization goal of using the above-mentioned model to optimize the operating variables to reduce the RON loss by more than 30% , meanwhile the sulfur content stayed relatively stable which was less than 5 ppm [47, 48].

Considering that the optimization of multiple variables often expands considerable computational effort, and the robustness of the sulfur content prediction model is remarkable, a simple random walk variable iteration method is proposed to iterate the variables [49]. In the iterative process of all manipulated variables, the distance between the variable and boundary condition is determined. If the variable is too close to the boundary, then the iterative direction must be opposite to the boundary. Otherwise, the variable obtains any positive or negative random iteration direction. The iteration results are substituted into the trained neural network model to obtain the sulfur content and octane number. If the optimization goal is met, or the number of iterations reaches the defined limit, then the optimization ends, and the iteration stops. If the optimization goal is not met, then the iteration is returned. Consequently, a solution to this problem is formed. The calculation and iteration processes are shown in Figure 7.

3.1 Relationship curve between each operating variable and the results based on No. 167 data

Figure 8 shows the influence of some operating variables on the change trend of sulfur content and RON value during the adjustment process within the operating range. The influence of operating variables on the RON value and sulfur content is a strong non-linear relationship. Changing the value of the operating variables during production according to the change trend of the model will have an impact on the product sulfur content and RON value.

As shown in the upper left picture, the sulfur content increases, and the octane number decreases with the change of variables. In the upper right picture, sulfur content and octane number increase. In the lower right picture, the sulfur content decreases, and the octane number increases. The complex influence of a single variable on the result is proven.

3.2 Visualization of the optimization process

Data No. 133 and No. 285 are randomly selected as examples, and 22 operating variables are iterated to the optimization target according to the iteration rules. The specific numerical changes are shown in Table 4. The sulfur content increases slightly after optimization but still less than 5 ppm. The RON loss is reduced from 1.3 to 0.8 and 1.1 to 0.7, which is reduced by 38% and 36%, and the optimization goal is achieved. Figure 9 and 10 shows that the sulfur content and octane number of the two samples change with some variables. Each variable randomly increases or decreases the iteration step within the range of the variable, and the model result will change accordingly.

Based on the above-mentioned ideas, the operating variables are optimized, and most of the samples can reach the optimization goal after iterations. Eighteen groups in 325 samples cannot reduce octane loss by more than 30%, which may be due to the nature of the raw materials and adsorbent. In general, the model established in this paper can complete the optimal design.

4. Result And Discussion

The algorithm first selects 30 principal components from 362 original data variables through PCA and uses the principal components as the initial clustering center of the K-means clustering algorithm to cluster the original variables. The variables closest to the cluster center are selected as the 30 key variables. Among them, the number of raw material property variables and catalyst property variables is only reduced by 25%, and the number of operating variables is reduced by 90%, indicating that raw material properties and catalyst properties contain more information than the original data variables in the desulfurization operating variables, which have a great impact on the results.

Subsequently, the neural network prediction models of the product sulfur content and octane number prediction model are determined. The average accuracy rate of the model prediction is 94% when the error value is within 0.15, and the average accuracy rate is 99% when the error value is within 0.3. This result indicates that the average accuracy rate of the model can remarkably predict the sulfur content and octane content of desulfurized gasoline.

Finally, using the above-mentioned model, an optimization algorithm called random walk iteration of operating variables is proposed, which can reduce the octane loss by more than 30% under the premise that the product sulfur content is less than 5 ppm. Using this algorithm on the No. 133 data sample, the optimized operating variables result in the increase of the sulfur content from 3.2 to 4.02 and reduction of the octane loss

from 1.3 to 0.8, attributing to a 38% reduction. Using this algorithm on the No. 285 data sample, the optimized operating variables result in the increase of the sulfur content from 4.03 to 4.37 and reduction of the octane loss from 1.1 to 0.7, attributing to a 36% reduction.

5. Conclusion

A PCA–KMeans backpropagation neural network prediction model that integrated PCA and K-means clustering algorithm was established to predict octane loss and sulfur content in the S ZORB SRT. The model has high prediction accuracy and can be used to reduce the optimization problem of octane loss. Then, the random walk iteration algorithm is proposed to reduce the octane loss by more than 30%, during desulfurization in the S ZORB SRT Reactor, meanwhile the sulfur content stayed relatively stable which was less than 5 ppm. The results of the proposed model are reliable and could be applied into the real industrialization, which are beneficial with both the efficiency of economy and environmental protection. Besides, the model is also suitable for other complex application scenarios with multiple variables, nonlinearities, and strong coupling in the chemical process, providing new ideas for model establishment and target optimization in the chemical process.

Declarations

- **Availability of data and materials**

The original data is submitted as an attachment.

- **Competing interests**

No competing interests.

- **Funding**

No funding.

- **Authors' contributions**

The first author Xiaoyi Geng completed most of the experiments and paper. Two correspondents Dr. Xin Wang and Dr. Guangcheng Zhang provided support for the theoretical basis and design of the PCA-Kmeans algorithm. The second author Bingyan Song and the third author Yu Chen wrote part of the paper and summarize the experiment.

- **Acknowledgements**

No acknowledgements.

References

[1] Nicodem, D.E., et al., *Photochemical processes and the environmental impact of petroleum spills.* Biogeochemistry, 1997. **39**(2): p. 121-138.^

[2] Patterson, D.J. and N.A. Henein, *Emissions from combustion engines and their control.* 1981.^

[3] Lyu, Y., et al., *Reactivation of spent S-Zorb adsorbents for gasoline desulfurization.* Chemical Engineering Journal, 2019. **374**: p. 1109-1117.^

[4] Montiel, C., R. Quintero, and J. Aburto, *Petroleum biotechnology: technology trends for the future.* African Journal of Biotechnology, 2009. **8**(12).^

[5] Breysse, M., et al., *Deep desulfurization: reactions, catalysts and technological challenges.* Catalysis Today, 2003. **84**(3): p. 129-138.^

[6] Hernández-Maldonado, A.J. and R.T. Yang, *New sorbents for desulfurization of diesel fuels via π-complexation.* AIChE Journal, 2004. **50**(4): p. 791-801.^

[7] Ali, S.A. and N.A. Al-Baghli, *Overview of FCC Gasoline Post-Treating Technologies*.^

[8] Wang, S.Q., et al., *Deep desulfurization of transportation fuels by characteristic reaction resided in adsorbents.* AIChE journal, 2009. **55**(7): p. 1872-1881.^

[9] Wang, X., et al. *Virtual Manufacturing System for Refinery Process*. in *2018 AIChE Annual Meeting*. 2018. AIChE.^

[10] XU, Y., *Advance in China fluid catalytic cracking (FCC) process.* Scientia Sinica Chimica, 2014. **44**(1): p. 13-24.^

[11] Sadare, O.O. and M.O. Daramola, *Adsorptive desulfurization of dibenzothiophene (DBT) in model petroleum distillate using functionalized carbon nanotubes.* Environmental Science and Pollution Research, 2019. **26**(32): p. 32746-32758.^

[12] Xiao, J., et al., *Reduction and Desulfurization of Petroleum Coke in Ammonia and Their Thermodynamics.* Energy & Fuels, 2016. **30**(4): p. 3385-3391.^

[13] Kong, Y., et al., *FCC gasoline desulfurization by pervaporation: Effects of gasoline components.* Journal of Membrane Science, 2007. **293**(1): p. 36-43.^

[14] Adžamić, T., K. Sertić-Bionda, and N. Marcec-Rahelic, *Modeling of the FCC Gasoline Desulfurization Process by Liquid Extraction with Sulfolane.* Petroleum Science and Technology, 2010. **28**(18): p. 1936-1945.^

[15] Bezverkhyy I, R.A., Gadacz G, et al., *Kinetics of thiophene reactive adsorption on Ni/SiO2 and Ni/ZnO* Catalysis Today, 2008. **130**(1): p. 199-205.^

[16] Schmidt, R. and E.L. Sughrue, *NH3 and HCl impact on sulfur removal from E-Gas™ gasification streams using S Zorb™ Gen. IV.* Fuel Processing Technology, 2010. **91**(6): p. 582-590.^

[17] Qiu, L., K. Zou, and G. Xu, *Investigation on the sulfur state and phase transformation of spent and regenerated S zorb sorbents using XPS and XRD.* Applied Surface Science, 2013. **266**: p. 230-234.^

[18] Jia Huayu, D.W., fan Chen, Yang Minglei, *Mechanism modeling of s-zorb reactor and parameter estimation based on improved whale swarm algorithm.* Journal of Chemical Engineering of Chinese Universities, 2018. **32**(6): p. 1395-1420.^

[19] Coley, Connor W., et al., *A graph-convolutional neural network model for the prediction of chemical reactivity.* Chemical Science, 2019. **10**(2): p. 370-377.^

[20] Gómez-Bombarelli, R., et al., *Automatic chemical design using a data-driven continuous representation of molecules.* ACS central science, 2018. **4**(2): p. 268-276.^

[21] Zhang, W., et al., *Predicting potential drug-drug interactions by integrating chemical, biological, phenotypic and network data.* BMC bioinformatics, 2017. **18**(1): p. 18.^

[22] Chiang, L., B. Lu, and I. Castillo, *Big data analytics in chemical engineering.* Annual review of chemical and biomolecular engineering, 2017. **8**: p. 63-85.^

[23] Statheropoulos, M., N. Vassiliadis, and A. Pappa, *Principal component and canonical correlation analysis for examining air pollution and meteorological data.* Atmospheric Environment, 1998. **32**(6): p. 1087-1095.^

[24] Saeed, F., et al., *Using graph-based consensus clustering for combining K-means clustering of heterogeneous chemical structures.* Journal of Cheminformatics, 2013. **5**(1): p. 1-3.^

[25] Litani-Barzilai, I., et al., *On-line remote prediction of gasoline properties by combined optical methods.* Analytica Chimica Acta, 1997. **339**(1): p. 193-199.^

[26] Wang, S., et al., *Feasibility study on prediction of gasoline octane number using NIR spectroscopy combined with manifold learning and neural network.* Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 2020. **228**: p. 117836.^

[27] Wang, S., et al., *Feasibility study on prediction of gasoline octane number using NIR spectroscopy combined with manifold learning and neural network.* Spectrochim Acta A Mol Biomol Spectrosc, 2020. **228**: p. 117836.^

[28] He, K., et al., *Near-infrared spectroscopy for the concurrent quality prediction and status monitoring of gasoline blending.* Control Engineering Practice, 2020. **101**: p. 104478.^

[29] Nan-ni, X., L. Zheng, and W. Peng, *Optimization of adsorptive desulfurization process conditions for FCC gasoline.* IOP Conference Series: Earth and Environmental Science, 2019. **267**: p. 022037.^

[30] Krishnan, R., V.A. Samaranayake, and S. Jagannathan, *A Multi-Step Nonlinear Dimension-Reduction Approach with Applications to Big Data.* IEEE Transactions on Knowledge and Data Engineering, 2019. **31**(12): p. 2249-2261.^

[31] Ding, C. and X. He, *Kmeans clustering via principal component analysis*, in *Proceedings of the twenty-first international conference on Machine learning*. 2004, Association for Computing Machinery: Banff, Alberta, Canada. p. 29.^

[32] Effendi, Y., et al., *Dimensionality Reduction using PCA and K-Means Clustering for Breast Cancer Prediction.* Lontar Komputer : Jurnal Ilmiah Teknologi Informasi, 2018.^

[33] Virupaksha, B., et al., *Analysis of naphthoquinone derivatives as topoisomerase I inhibitors using fragment based QSAR.* Journal of Cheminformatics, 2013. **5**(S1): p. P22.^

[34] Wold, S., K. Esbensen, and P. Geladi, *Principal component analysis.* Chemometrics and intelligent laboratory systems, 1987. **2**(1-3): p. 37-52.^

[35] Ivosev, G., L. Burton, and R. Bonner, *Dimensionality reduction and visualization in principal component analysis.* Analytical chemistry, 2008. **80**(13): p. 4933-4944.^

[36] Vasan, K.K. and B. Surendiran, *Dimensionality reduction using Principal Component Analysis for network intrusion detection.* Perspectives in Science, 2016. **8**: p. 510-512.^

[37] Lever, J., M. Krzywinski, and N. Altman, *Points of significance: Principal component analysis*. 2017, Nature Publishing Group.^

[38] Boutsidis, C., et al., *Randomized dimensionality reduction for $ k $-means clustering.* IEEE Transactions on Information Theory, 2014. **61**(2): p. 1045-1062.^

[39] Cohen, M.B., et al. *Dimensionality reduction for k-means clustering and low rank approximation*. in *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*. 2015.^

[40] Lafon, S. and A.B. Lee, *Diffusion maps and coarse-graining: A unified framework for dimensionality reduction, graph partitioning, and data set parameterization.* IEEE transactions on pattern analysis and machine intelligence, 2006. **28**(9): p. 1393-1403.^

[41] Gan, G. and M.K.-P. Ng, *k-means clustering with outlier removal.* Pattern Recognition Letters, 2017. **90**: p. 8-14.^

[42] Cheriyadat, A. and L.M. Bruce. *Why principal component analysis is not an appropriate feature extraction method for hyperspectral data*. in *IGARSS 2003. 2003 IEEE International Geoscience and Remote Sensing Symposium. Proceedings (IEEE Cat. No. 03CH37477)*. 2003. IEEE.^

[43] Benesty, J., et al., *Pearson correlation coefficient*, in *Noise reduction in speech processing*. 2009, Springer. p. 1-4.^

[44] Chua, L.O. and L. Yang, *Cellular neural networks: Theory.* IEEE Transactions on circuits and systems, 1988. **35**(10): p. 1257-1272.^

[45] Yegnanarayana, B., *Artificial neural networks*. 2009: PHI Learning Pvt. Ltd.^

[46] Tang, B., et al., *A self-attention based message passing neural network for predicting molecular lipophilicity and aqueous solubility.* Journal of Cheminformatics, 2020. **12**(1): p. 1-9.^

[47] Del Castillo, E., *Process optimization: a statistical approach*. Vol. 105. 2007: Springer Science & Business Media.^

[48] Motlaghi, S., F. Jalali, and M.N. Ahmadabadi, *An expert system design for a crude oil distillation column with the neural networks model and the process optimization using genetic algorithm framework.* Expert systems with applications, 2008. **35**(4): p. 1540-1545.^

[49] Chan, L., G.R. Hutchison, and G.M. Morris, *Bayesian optimization for conformer generation.* Journal of cheminformatics, 2019. **11**(1): p. 1-11.^

Tables

TABLE 1 Cumulative contribution rate of main components

	Number of variables	Principal components	Cumulative contribution rate
Raw material properties	7	5	≥0.9
Adsorbent properties	4	3	≥0.9
Operating variables	351	22	≥0.85
Sum	362	30	0.84

TABLE 2 Results of variables clustering

	Number of variables	Number of clusters	Variable name (partial)
Raw material properties	7	5	Olefin, Density, Saturated hydrocarbon, Bromine, RON
Adsorbent properties	4	3	Spent S, Spent C, Regenerated S
Operating variables	346	22	S ZORB.TE_7508.DACA, S ZORB.FT_3301.TOTAL, S ZORB.PT_7508B.DACA, S ZORB.PT_2301.PV, S ZORB.TC_1606.PV, S ZORB.CAL.SPEED. PV, S ZORB.TE_1106.DACA. PV, S ZORB.PC_2401.PIDA. OP, S ZORB.SIS_TE_2605.PV, S ZORB.FT_9301.PV...

TABLE 3 Prediction results of PCA-KMeans-BP neural network

Predictor variable	Error from true value	Accuracy	Error from true value	Accuracy
Sulfur content	≥0.15	93%	<0.3	100%
RON	≥0.15	95%	<0.3	98%

TABLE 4 Partial operating variable range table

	SZORB.PC_2401B.DACA	SZORB.PT_7103.DACA	SZORB.TE_7508.DACA	SZORB.PT_7502.DACA	SZORB.FT_5104.TOTAL
Min	0.05	-0.5	-2000	20	12
Max	2.35	4000	45	30	14000
Step	0.1	100	1	1	500

TABLE 5 Optimization process of NO.133 and NO.285 samples

Number	Optimization stage	Raw material sulfur content/ppm	Raw material RON	Product sulfur content/ppm	Product RON
133	before optimization	248.0	89.4	3.2	88.09
	after optimization	248.0	89.4	4.02	88.6
285	before optimization	199.0	89.3	4.03	88.18
	after optimization	199.0	89.3	4.37	88.6

Figures

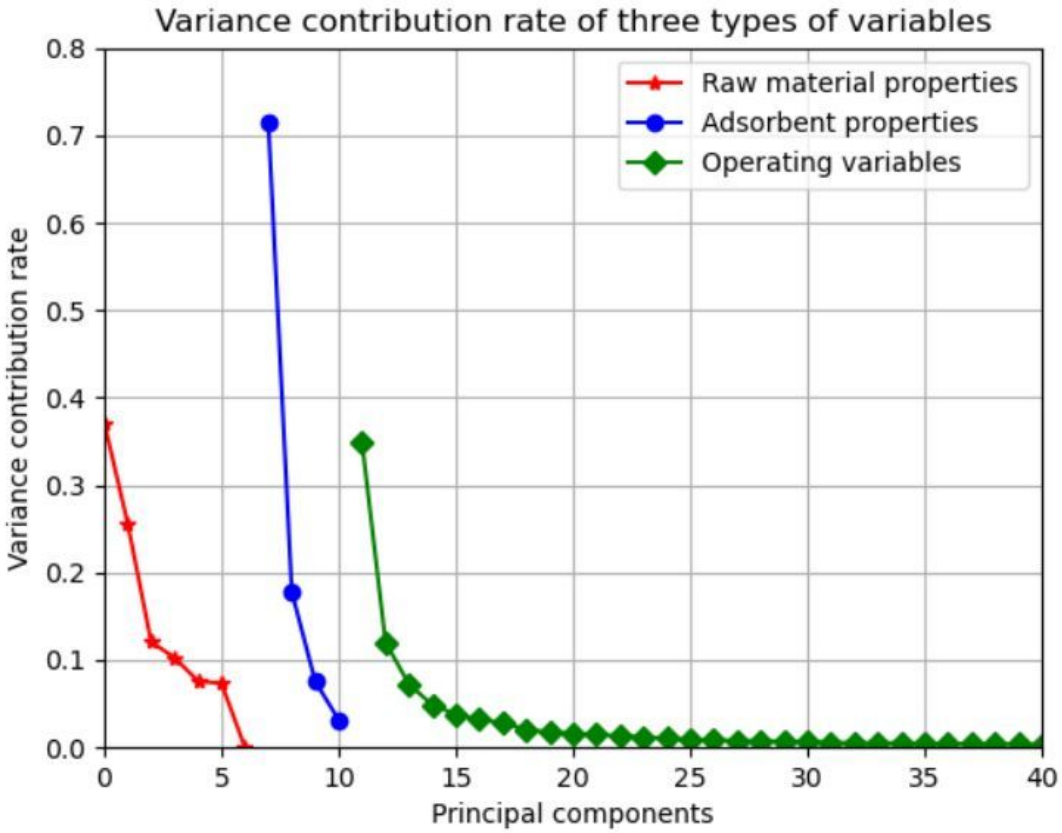
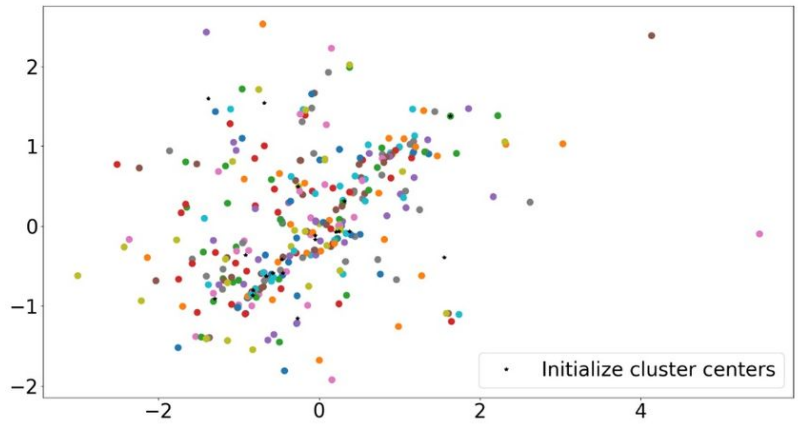


Figure 1

The results of PCA of three kinds of original variables

(a) Initial distribution of 346 operating variables and initialization of cluster centers based on PCA



(b) Operating variables clustering results

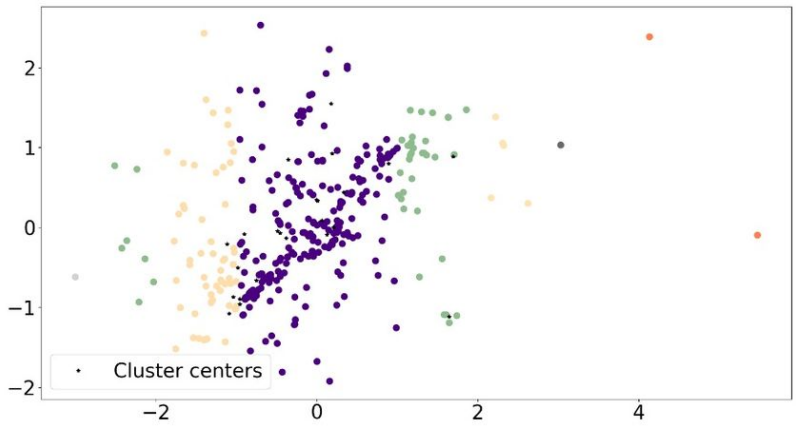


Figure 2

Cluster procedure of operating variables

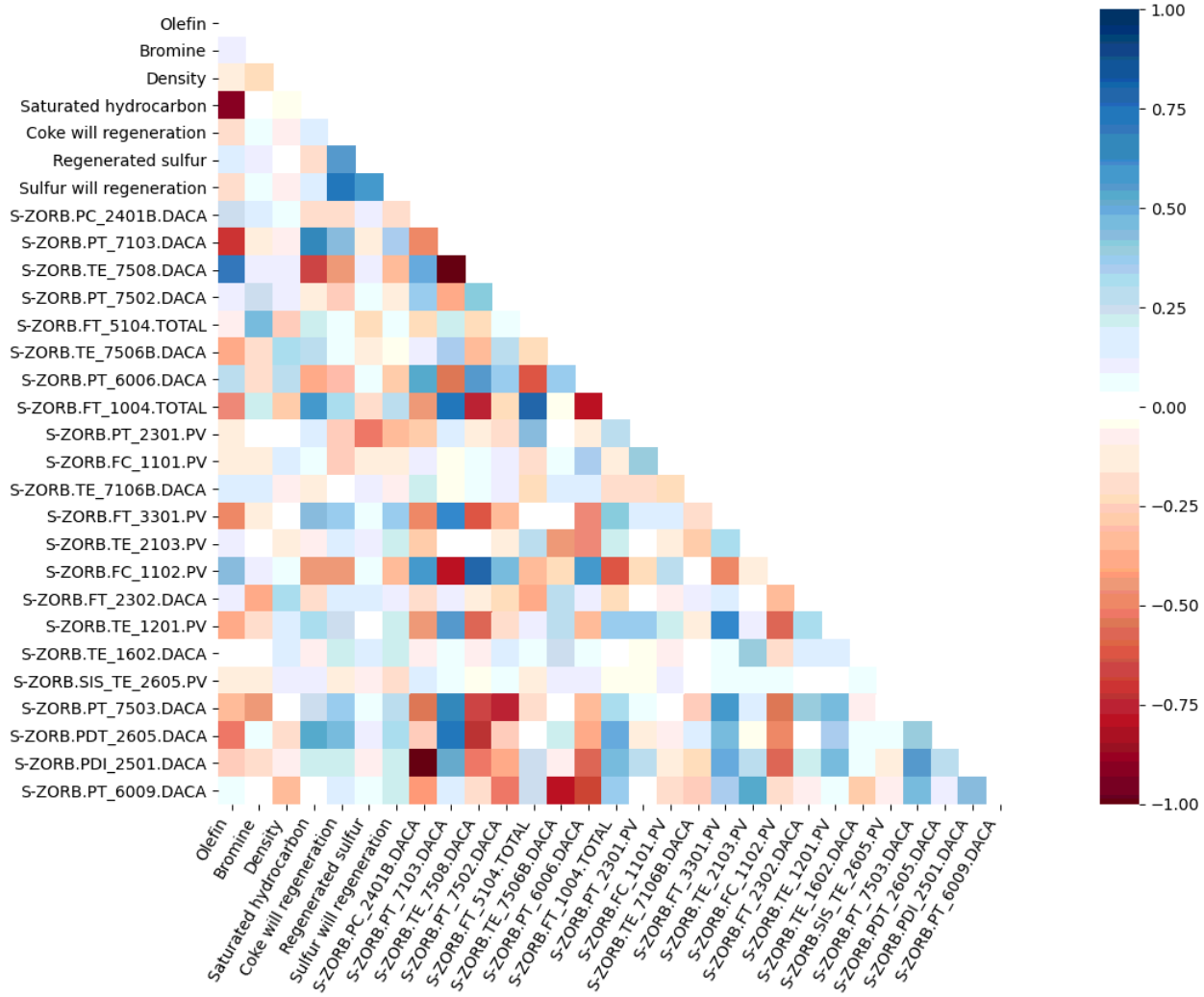


Figure 3

Correlation Analysis of 30 key variables

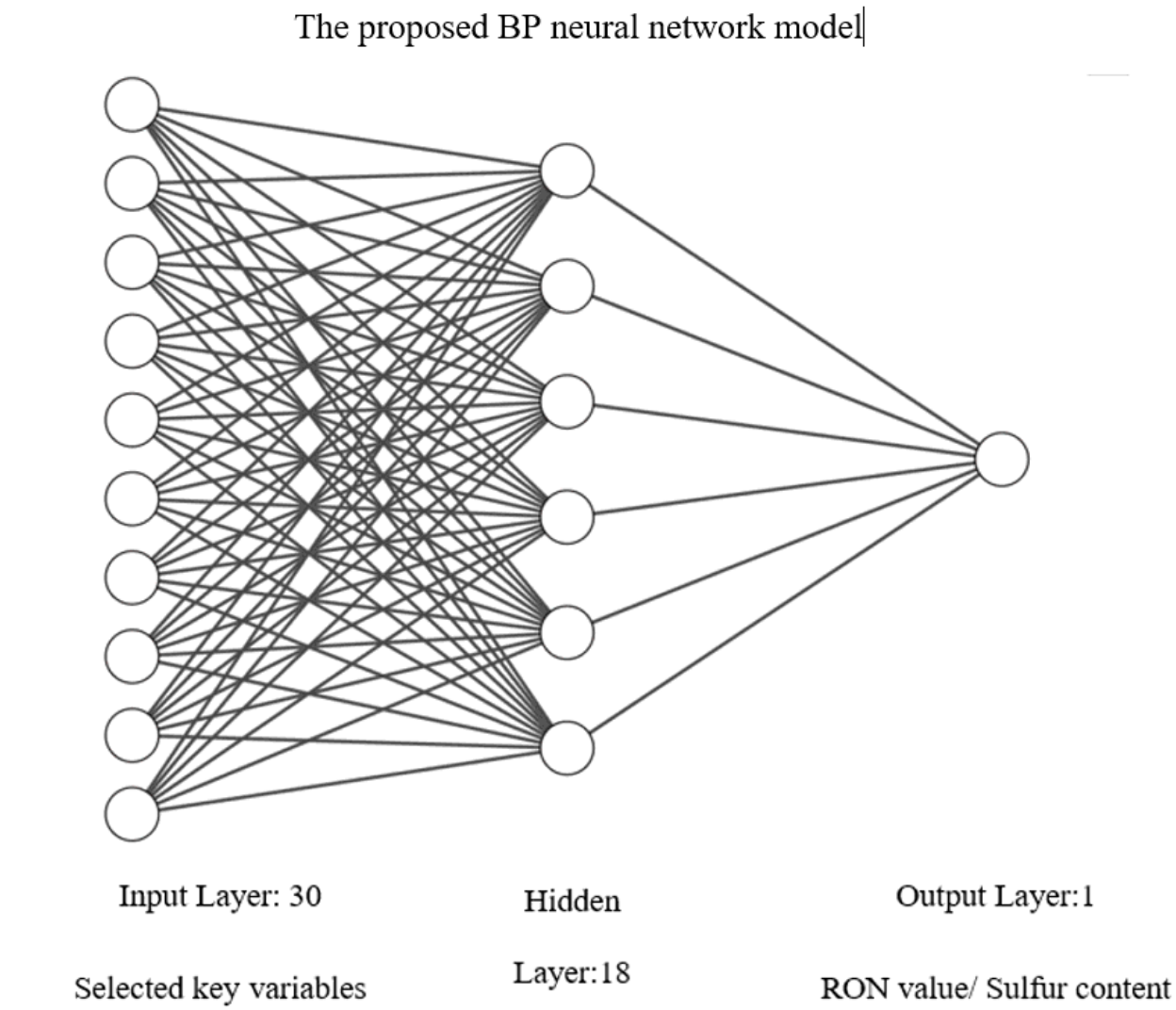


Figure 4

The proposed BP neural network model

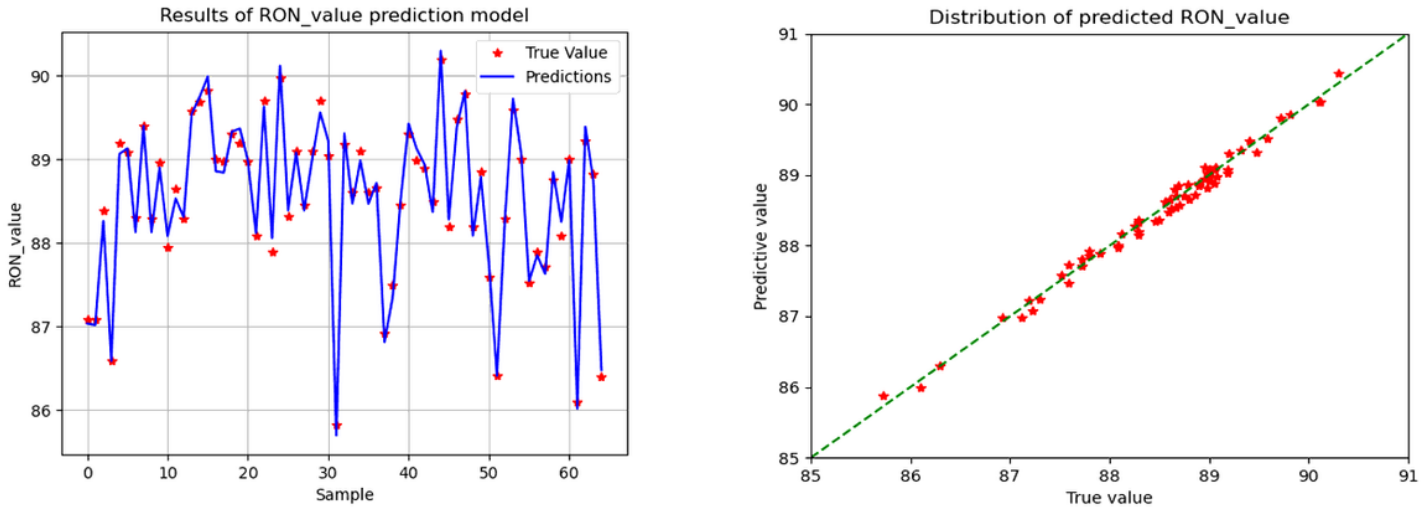


Figure 5

Comparison of predicted data with measured data of RON_value

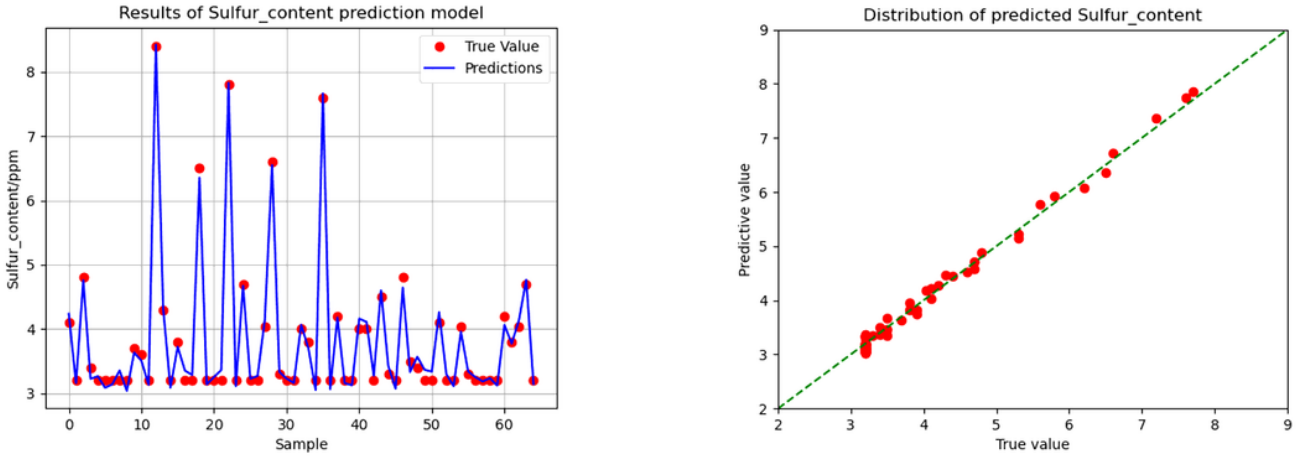


Figure 6

Comparison of predicted data with measured data of Sulfur_content

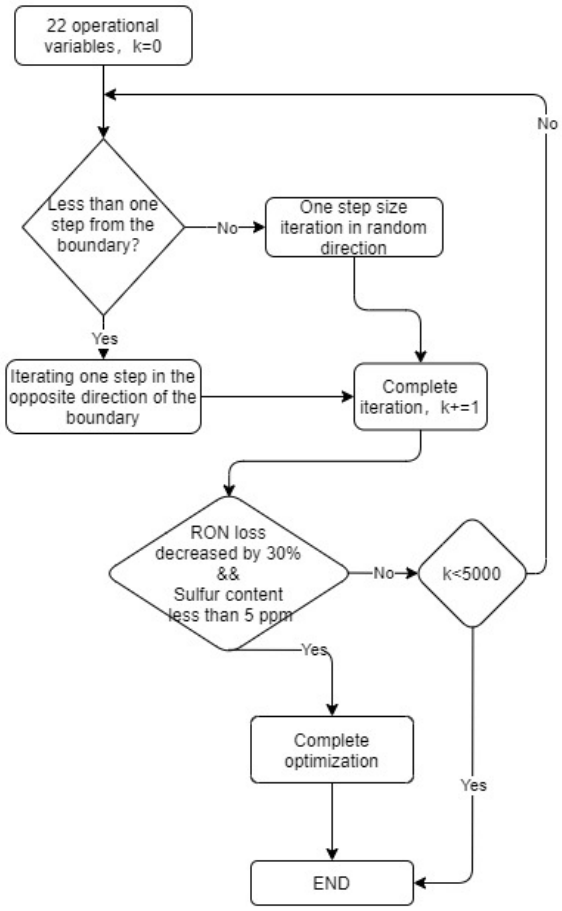


Figure 7

Optimization flow chart

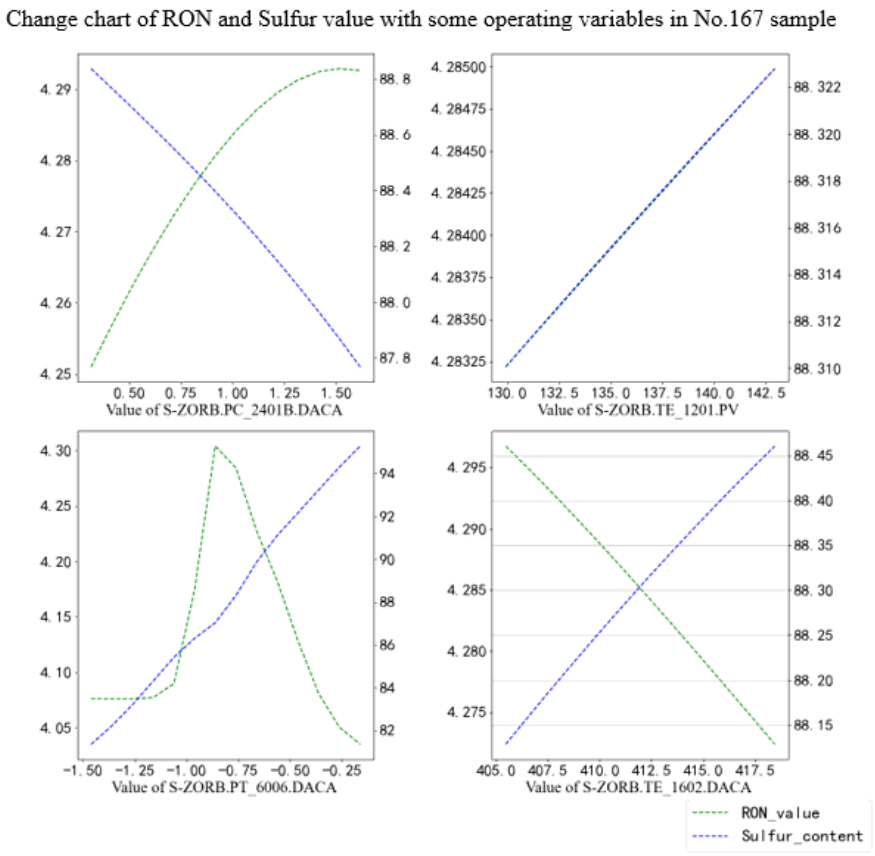


Figure 8

Change trend of product with some operating variables

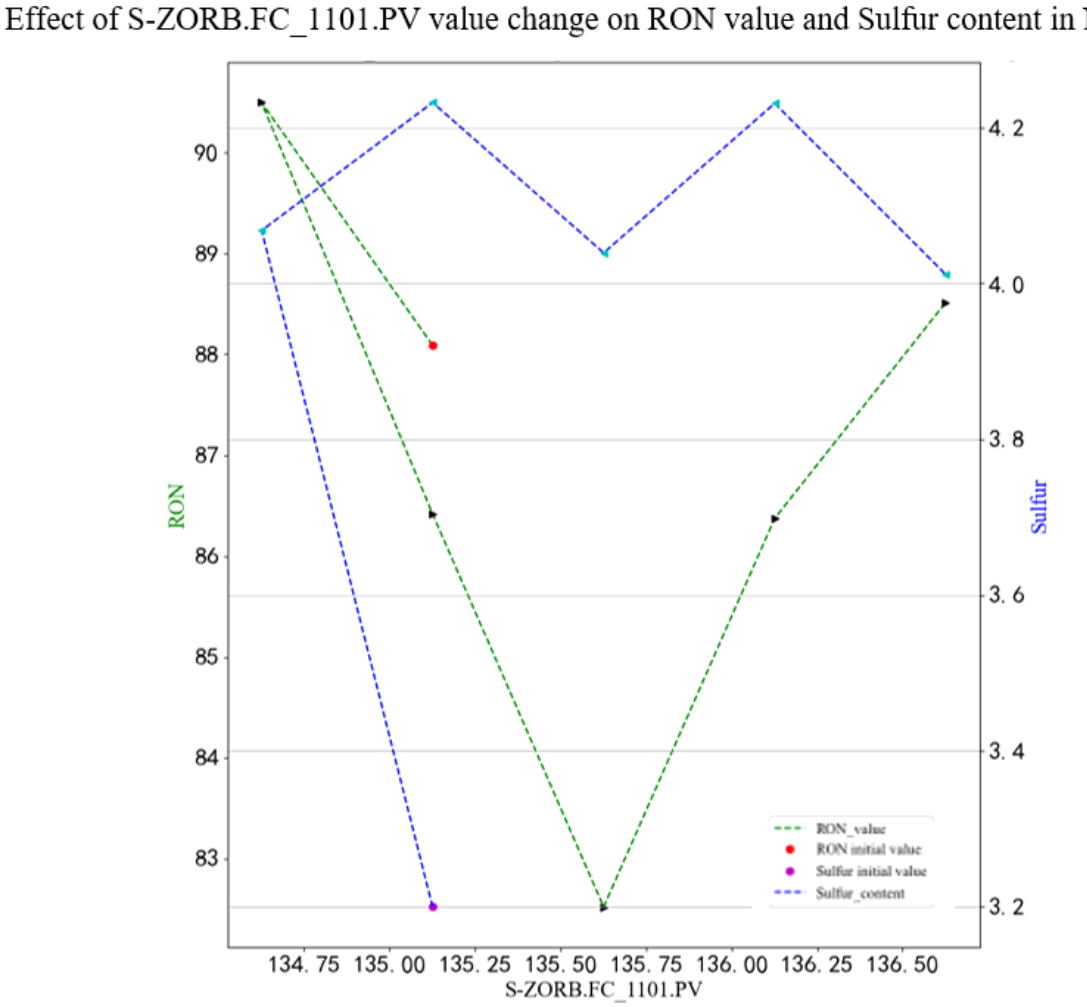


Figure 9

Visualization of some variables of sample 133 in optimization process

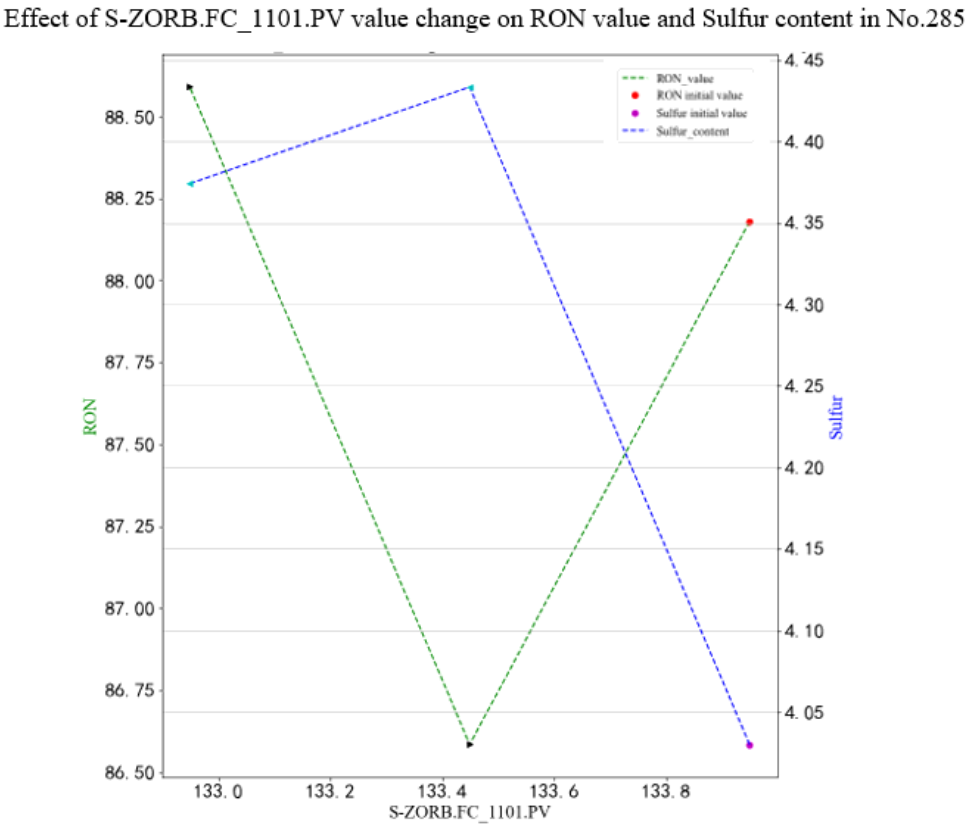


Figure 10

Visualization of some variables of sample 285 in optimization process

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Sampledata.xlsx