Multicollinearity Resolution Based on Machine Learning: A Case Study of Carbon Emissions

Xuanming Zhang [1]

Abstract

This study proposes a novel analytical framework that integrates DBSCAN clustering with the Elastic Net regression model to address multifactorial problems characterized by structural complexity and multicollinearity, exemplified by carbon emissions analysis. DBSCAN is employed for unsupervised learning to objectively cluster features, while the Elastic Net is utilized for high-dimensional feature selection and complexity control. The Elastic Net is specifically chosen for its ability to balance feature selection and regularization by combining L1 (lasso) and L2 (ridge) penalties, making it particularly suited for datasets with correlated predictors. Applying this framework to energy consumption data from 46 industries in China (2000-2019) resulted in the identification of 16 categories. Emission characteristics and drivers were quantitatively assessed for each category, demonstrating the framework's capacity to identify primary emission sources and provide actionable insights. This research underscores the global applicability of the framework for analyzing complex regional challenges, such as carbon emissions, and highlights qualitative features that humans find meaningful may not be accurate for the model.

1. Introduction

Green and low-carbon development has become a crucial strategy for countries worldwide to promote economic sustainability. A recent IPCC report has once again emphasized that human induced greenhouse gas emissions are the primary driver of global warming. As the largest emitter of carbon dioxide, China faces significant challenges in achieving its carbon neutrality goals (Luo et al., 2024).

China plays a pivotal role in the world economically. Research shows that China is a major energy producer, with coal accounting for 11.6% of international output in 2019. In 2020, coal dominated 66.6% energy mix (He et al., 2025). However, accelerated industrialization and urbanization pose low-carbon transition challenges as coal dependency and emissions are significant. As a primary coal and industrial base, energy-industrial restructuring in China will enormously impact reductions worldwide. Understanding key industry emissions in China is critical to inform reduction policies.

When investigating carbon emissions in China, we observed that direct estimation and assessment of carbon emissions were often challenging. This difficulty arises from dealing with data featuring multiple types and structural characteristics, such as the complex relationships within energy consumption patterns among different industries.

In related studies, classical statistical models such as STIRPAT models have difficulties in effectively distinguishing the relationships between variables when facing serious multicollinearity problems, and cannot deeply mine subtle features from data (Li et al., 2024). Divisia decomposition method and logarithmic average decomposition method lack uniqueness in results and cannot fundamentally solve the problems (Xie et al., 2016). Although deep learning methods such as neural networks can approximate arbitrary complex functions to a certain extent, they require a large amount of high-quality data support. Carbon emission data are often limited in quantity and noisy, with low efficiency of information transmission between layers, making it difficult to comprehensively capture industry characteristics by spanning across complex factor relationships (Feda et al., 2024).

In order to better solve such multi-factor problems and promote generalized applications, we believe that the essence lies in the serious multicollinearity problems that exist in feature mining and processing multi-source data. And we propose an algorithm framework based on DBSCAN clustering and penalized regression, aiming to achieve the following objectives:

- DBSCAN clustering can identify clustering results

[1] School of Computer, Data & Information Sciences, University of Wisconsin-Madison, Madison, USA. Correspondence to: Xuanming Zhang <xzhang2846@wisc.edu>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025. Copyright 2025 by the author(s).

based on density differences in an unsupervised learning condition, without needing to pre-define the number of categories, which can better reflect the intrinsic patterns in data.
- Penalized regression can deeply mine the important relationships between features in models through controlling model complexity and high-dimensional feature selection, complementing the interpretability deficiency of deep learning methods.
- This DBSCAN-Penalized regression (DPR) framework adopts a hierarchical processing approach of first clustering and then regression, effectively solving the problem of multicollinearity while also considering the degree of model complexity, with better fitting effects.

For this purpose, this paper refers to the theories of DBSCAN and penalized regression (Hoerl & Kennard, 2000), proposing a general analytic framework DPR that combines data feature clustering and data regression. DBSCAN can form clusters of arbitrary shapes (Ullmann et al., 2022), and penalized regression realizes sparse feature selection and controls associativity through L1/L2 regularization (Mozafari-Majd & Koivunen, 2023). It is successfully applied to carbon emission data of 46 industries in China, aiming to compensate for the shortcomings of traditional models and deep learning in this area. Empirical results show that this framework can identify major emission sources and driving factors to provide quantitative references for decision making.

2. Methodology

This study uses energy consumption and carbon emissions data from 46 different industries in China spanning from 2000 to 2019 for model training, as a case study to validate the effectiveness of the DPR framework in addressing multifactor collinearity issues. Specifically, we employ the DBSCAN algorithm to cluster industries, followed by the application of penalized regression methods to quantitatively analyze the relationships between influencing factors and emissions, revealing the characteristic emission profiles and driving factors for each industry.

These methods unveil the characteristic emission profiles and dominant driving mechanisms within different industry categories. Appendix A.1 and Appendix A.2 will provide detailed introductions to DPR techniques, while Section 3 will present preliminary clustering analysis results.

The results from the real-world case study demonstrate that DPR framework effectively identifies major emission sources and resolves multicollinearity issues within the data, providing quantitative foundations for decision-making. The primary objective of this study is to validate the effectiveness of the clustering-regression algorithm framework in addressing multifactor issues, the overview of our method is shown in Fig 1.

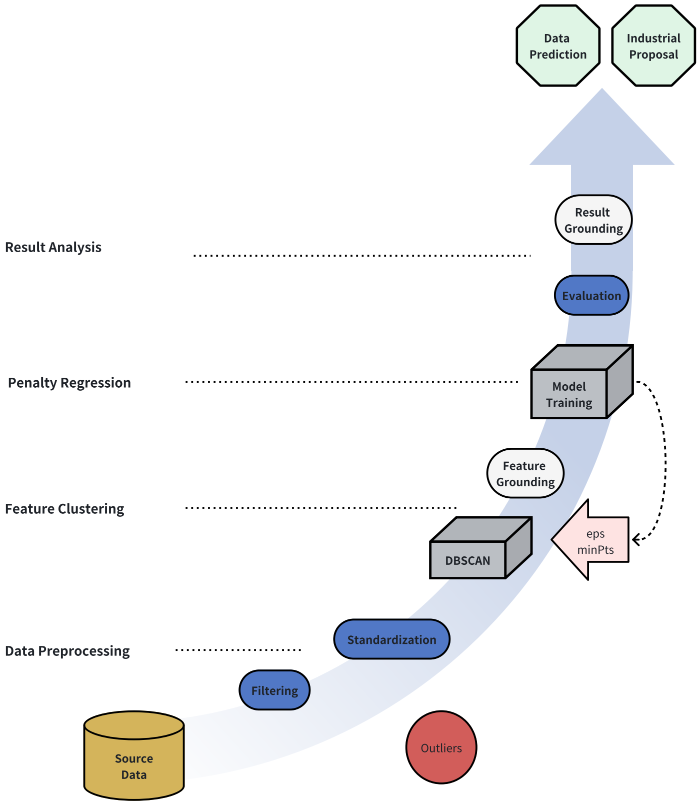

Figure 1. DPR Workflow.

3. Experiment

3.1. Data Sources

The sources of CO2 emissions data in this paper include the China Carbon Accounting Database and the Statistical Yearbook. Among them, Carbon Emission Accounts & Datasets (CEADs, 2020) provides carbon dioxide emissions data from 2000 to 2019 (Zhu & Li, 2024), and the sources of the data include internet surveys, national and local government statistics, etc. In contrast, the National Statistical Yearbook provides relevant economic and energy statistics for provinces, including data on industry, agriculture, transportation, energy and environmental protection. Carbon Emission Accounts & Datasets (CEADs, 2020), for example, covers 17 energy types, as summarized in Table 1.

3.2. Pre-Processing

In data pre-processing, 16 petroleum products with zero energy consumption values in China were removed, leaving

16 effective energy sources. Logging and transportation of wood/bamboo with zero values each year were also excluded, giving 46 effective industries.

Number	Energy Type	Number	Energy Type
1	Raw Coal	10	Gasoline
2	Cleaned Coal	11	Kerosene
3	Washed Coal	12	Diesel Oil
4	Briquettes	13	Fuel Oil
5	Coke	14	LPG
6	Coke Oven Gas	15	Refinery Gas
7	Other Gas	16	Petroleum Products
8	Other Coking Products	17	Natural Gas
9	Crude Oil		

Table 1. *Energy types in dataset*

To initially explore differences in energy types, energy use from 2012-2019 was visualized in Fig 2. This revealed variations that informed subsequent clustering and modeling to characterize industries and identify reduction opportunities.

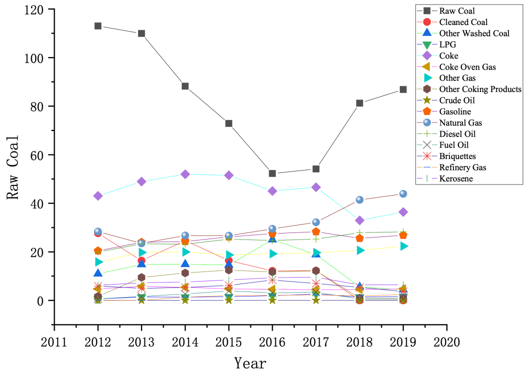

Figure 2. *Schematic representation of total CO2 emissions by energy source, 2012-2019, raw coal distribution.*

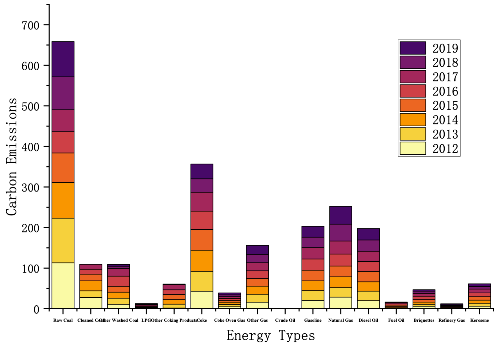

Figure 3. *Schematic representation of total CO2 emissions by energy source, 2012-2019, carbon emission distribution.*

Fig 3 demonstrates energy use trends over time. Raw coal consumption declined as natural gas rose, though coal remained dominant per the stacked bar graph of carbon emissions. Increasing gasoline, diesel and natural gas suggest rapid economic growth driving multi-sector energy demand.

The carbon emissions data in China covers 46 sectors, exhibiting variability over time. To facilitate comparisons across sectors, the data for each sector was also normalized using Equation:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

This initial exploration revealed industry energy characteristics and transitions informing subsequent clustering and modeling to systematically characterize emissions profiles and identify approaches supporting carbon reduction goals.

Carbon emissions data are concentrated, with huge differences in absolute values between different industries. Some industries have emissions as high as 39.03 Mt of coking coal, while the highest is only 0.045 Mt of gasoline compared to other industries. The maximum and minimum standardization better shows the fine differences and changing trends between the data. For example, the highest share of coking coal in ferrous metal smelting and the highest use of raw coal in food processing are better represented.

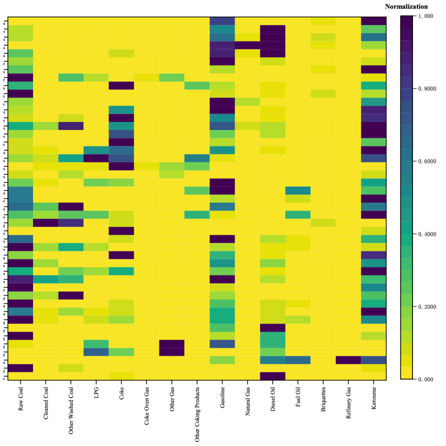

Figure 4. *Heat map of energy consumption by sector, normalized data.*

It is clear from Fig 4 that with the same number of classifications, the energy types in most industries are not well differentiated due to the influence of the absolute size of the original data, and the color blocks are basically dominated by yellow, whereas the comparison of the heat map after normalization in the graph on the right eliminates the masking effect caused by the deviating absolute data, and the energy levels in each industry are richer and finer, and the characteristics are more obvious.

Sector	Raw Coal		Cleaned Coal		Briquettes		Coke	
	Raw Data	Normalized Data	Raw Data	Normalized Data	Raw Data	Normalized Data	Raw Data	Normalized Data
Metal	0.552	0.014	2.460	0.063	0.037	0.047	39.03	1.000
Electricity	29.105	1.000	0.015	0.000	3.224	0.110	0.026	0.001
Food	0.482	1.000	0.025	0.051	0.036	0.075	0.067	0.140
Rubber	0.026	0.582	0.001	0.003	0.000	0.012	0.000	0.020
Furniture	0.011	0.170	0.003	0.001	0.000	0.000	0.067	1.000

Table 2. Comparison of Excerpt Normalized Data

3.3. DBSCAN Cluster Analysis

MATLAB optimization yielded optimal **SC**=0.6, **SSE**=5 at **C**=16 clusters, showing clear inter-cluster variation and high intra-cluster similarity. Validation using additional years and metrics supported robust clustering.

Given the unsupervised nature of DBSCAN, we further validated clustering robustness using additional normalized data from other years, along with confusion matrices and ROC curves, as shown in Fig 5 and 6

Figures show the quadratic SVM confusion matrix achieved an average true class accuracy of 93.7%, except for isolated classes. Most ROC curves converged closely to 1. Therefore, we believe the clustering yielded reasonable classification at the energy feature level, with samples exhibiting good divisibility and concentration within clusters. Clustering validation results are summarized in Table 7.

On the basis of this classification, we extracted data features for the different cluster classes and calculated the corresponding means, variances, medians and gave the common quartiles, shown in Table 3.

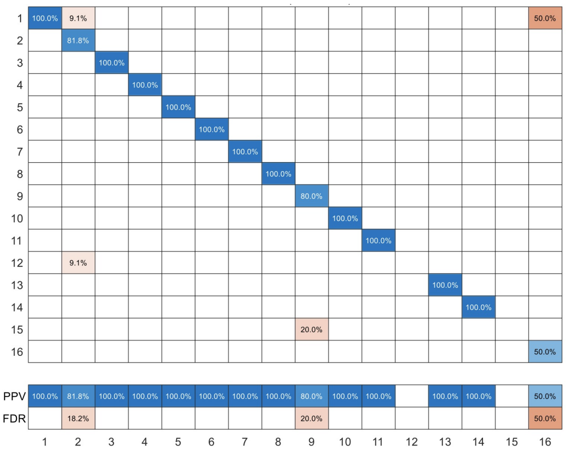

Figure 5. *SVM [7] - Quadratic Confusion Matrix.*

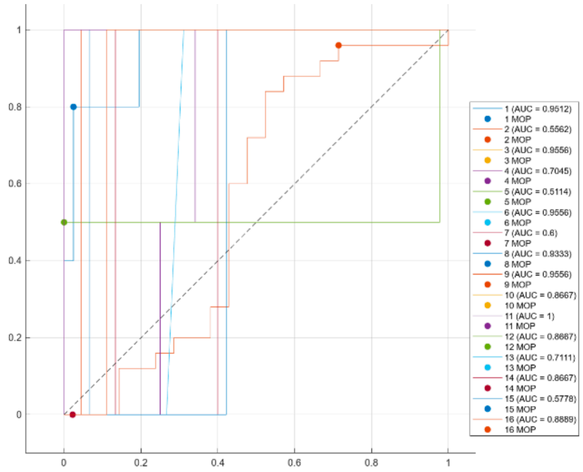

Figure 6. *SVM [7] - ROC.*

Class	Sum	Mean	Variance	Minimum	$P_{0.25}$	Median	$P_{0.75}$	Maximum
1	23.51	1.57	11.24	0	0	0.00	1.23	12.91
2	67.25	4.48	99.43	0.0048	0.198	0.63	3.13	40.41
3	5.77	0.38	0.37	0	0	0.00	1.20	1.85
4	3.60	0.24	0.22	0	0	0.02	0.33	1.64
5	1.08	0.07	0.02	0	0	0.00	0.15	0.43
6	3.86	0.26	0.20	0	0.002	0.03	0.24	1.63
7	0.07	0.00	0.00	0	0	0.00	0.01	0.04
8	24.17	1.61	11.72	0	0.004	0.01	0.41	10.96
9	66.22	4.41	99.29	0	0.049	1.05	3.48	40.56
10	18.13	1.21	6.36	0	0	0.05	0.64	8.11
11	10.59	0.71	0.91	0	0	0.36	1.06	3.76
12	0.52	0.03	0.00	0	0	0.00	0.06	0.17
13	20.15	1.34	3.06	0	0.012	0.45	1.87	4.51
14	2.62	0.17	0.05	0	0	0.03	0.42	0.69
15	1.48	0.10	0.02	0	0	0.03	0.21	0.39
16	27.60	1.84	11.56	0	0	0.00	3.00	8.99

Table 3. *Cluster characteristics*

The clustering results were visualized to obtain the energy consumption pile-up and box line diagrams of different clusters in Fig 7. It can be seen that the second and ninth clusters are comparable in terms of total carbon emissions, but with significant differences in energy structure. In the context of model regression, and based on our experience in handling carbon emissions data, we will apply a logarithmic transformation to the imported data in regression.

3.4. Regression Analysis

3.4.1. Training

In ridge regression, we introduce an L2 penalty term controlled by the regularization coefficient λ. A larger λ enforces stronger regularization, resulting in a simpler model with poorer training fit but better generalization. We iterate λ values to find the optimal trade-off, maximizing R^2 on validation data.

By iterating λ from 0 to 0.5 with a step size of 0.01, the maximum R^2 was 0.74691 at some λ value, but with a high MSE of 0.49342, indicating poor fitting ability. This and the dense solution (s=1) suggested ridge regression was not optimal for handling the high-dimensional data. Further regularization was needed.

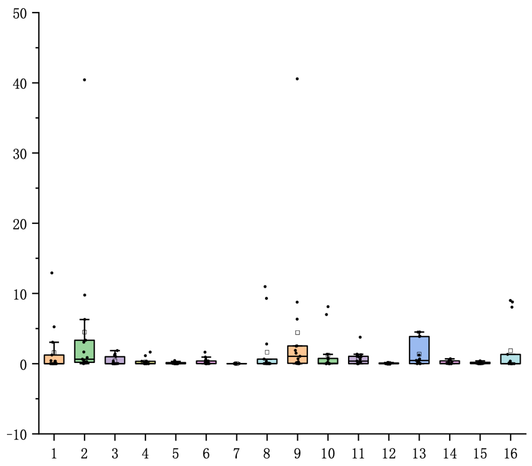

Figure 7. Energy distribution chart.

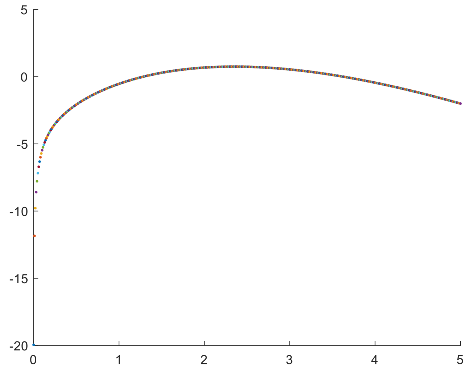

Figure 8. R^2 iteration chart.

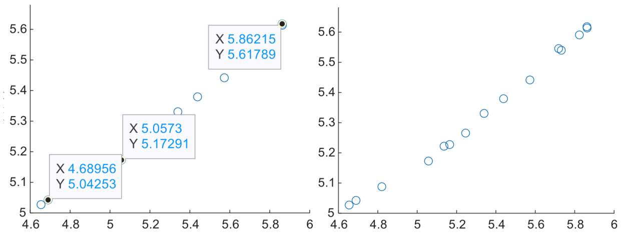

Figure 9. Ridge regression fitting effect chart.

Looking again at the predicted and actual values, using (5.61789,5.86215) as an example, we can see that $\exp(5.61789)$ is about 275.308 $MtCO_2$ and $\exp(5.86215)$ is about 351.479 $MtCO_2$, with a predicted difference of 76.17 $MtCO_2$, a deviation of 27.67%.

We applied Lasso regression to select influential factors by introducing an L1 regularization term λ. Lasso drives some coefficients to exact zero values through its L1 penalty, effectively reducing the dimensionality of data. This mitigates overfitting while selecting important covariates. We plotted the lasso coefficient trajectories against the regularization strength λ to identify the optimal λ value balancing prediction and interpretation.

The lasso coefficient trajectories illustrate how coefficients shrink with increasing λ regularization. The x-axis plots λ values, y-axis the coefficients. As λ grows, curves plateau horizontally, leaving preceding portions non-zero. Cross-validation determined the optimal λ with lowest error. The plot shows seven influential features retained, explaining 99.91% variation (R^2 =0.9991) with little error ($MSE = 1.4774 \times 10^{-4}$). Plugging nonzero coefficients into the regression equation $\ln I = \sum_{i=1}^{n=7} a_i \ln x_i + coef$ yields the influential factors as:

$$\ln I = 0.2823 \ln x_1 + 0.1203 \ln x_2 + 0.1660 \ln x_3 + 0.0200 \ln x_4 + 0.0850 \ln x_5 + 0.0005 \ln x_6 + 0.3461 \ln x_7 + 2.8986$$

Bringing in the data for solution and observing the fit effect, a plot of the fit effect of the actual observed and predicted values is obtained. Taking (5.0573,5.05189) as an example, the prediction difference is at 0.85 $MtCO_2$ and the error is only at 0.54%, which is a significant improvement over the ridge regression prediction.

However, Lasso regression resulted in a sparse solution with 7 nonzero coefficients and sparsity level s=0.4375, indicating potential for further dimensionality reduction. Therefore, we applied elastic net regression to balance L1 and L2 regularization, optimizing the α hyperparameter via cross-validation. The optimal model achieved near-perfect fit on validation data (R^2=0.9999, $MSE = 1.8309 \times 10{-}5$), as evidenced by its linear fitting abilities shown below. Elastic net proved superior for modeling carbon emissions with this high-dimensional yet sparse dataset.

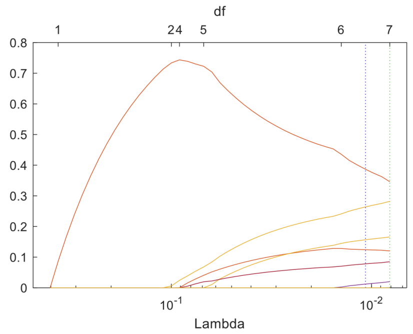

Figure 10. Lasso coefficient fitting trajectory chart.

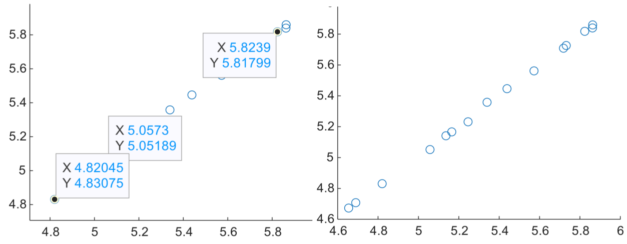

Figure 11. Lasso fitting effect chart.

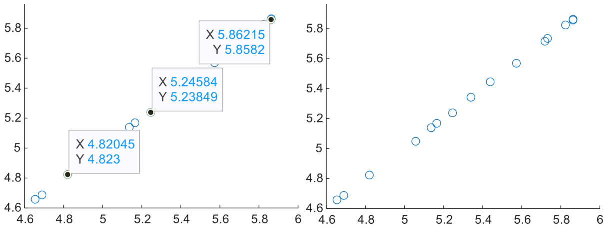

Figure 12. Elastic Net fitting effect chart.

3.4.2. EVALUATION

A combined evaluation of coefficients and evaluation indicators has been produced from the three regression results in Table 4.

Class	Ridge		Lasso		Elastic Net	
a_0	4.6178	$Coef0$	0	$Coef0$	0	$Coef0$
1	0.0390	-	0	2.8986	0.0236	2.3273
2	0.0628	R^2	0.2823	R^2	0.4319	R^2
3	0.0159	0.74691	0	0.9991	0.0669	0.9994
4	0.0139	Mse	0	Mse	0	Mse
5	0.0125	0.49342	0	1.4774×10^{-4}	0	1.8309×10^{-5}
6	0.0334	0.49342	0	1.4774×10^{-4}	0	1.8309×10^{-5}
7	0.0342	s	0	s	0.0108	s
8	0.0321	1	0.1203	0.4375	0.1132	0.75
9	0.0341	λ	0.1660	λ	0.1534	$\lambda_1 = \lambda_2$
10	0.0268	2.35	0.0200	0.0081	0.0540	2.7826×10^{-4}
11	0.0072	2.35	0	0.0081	0.0058	2.7826×10^{-4}
12	-0.0070	α	0	α	0	α
13	0.0256	-	0.850	-	0.0538	0.5
14	0.0292	-	0.0005	-	0.0797	0.5
15	0.0359	-	0.3461	-	0.0368	0.5
16	0.0344	-	0	-	0.0047	0.5

Table 4. Comprehensive table of coefficients and evaluation indicators

It can be seen that the second cluster has the largest proportion of the three regression coefficients, indicating a strong influence on carbon emissions.

The energy distribution Table 4 and pie chart in Fig 13 show that the second cluster of industries covers three sectors, including energy production, light manufacturing and high-tech industries, and that the energy supply is mainly based on raw coal.

Compared to other fossil fuels, raw coal has a low energy density and a low combustion efficiency of about 30%, and the raw coal obtained from coal mining and beneficiation is returned to industry for further development, while most of it is concentrated in the supply of electricity and hot water. The reasons for this are the relatively rapid economic development of China, coupled with the geographical constraints of its inland industrial location, complex terrain for industrial distribution, difficult transport and electricity transmission, which make the cities more dependent on coal resources for energy consumption, and the relatively high consumption of raw coal.

Due to the nature of the index, the ridge regression predictions are not very different after taking $\ln$, but the error becomes increasingly large over time for actual carbon emissions. While the lasso and elastic net still provide an excellent fit to actual carbon emissions, there is a partial bias in the lasso at the end of 2014.

Within a certain range of hyperparameters, ridge regression can improve the generalization and interpretability of the model to a certain extent, but the ridge regression method cannot directly eliminate features with small or zero weights and leads to feature scaling problems; LASSO regression can significantly improve the interpretability and sparsity of the model while maintaining high prediction accuracy, but the LASSO regression method ignores the weakly correlated features, but the LASSO regression method ignores the prob-

Socio-economic Sectors	Cluster	Industry Category	Major Energy Distribution	propotion
Coal Mining and Dressing	1	Energy Production	Raw Coal	22.49%
Nonmetal Minerals Mining and Dressing	1	Energy Production	Raw Coal	1.2%
Food Processing	1	Light Manufacturing	Raw Coal/Natural Gas	1.5%
Beverage Production	1	Light Manufacturing	Raw Coal	1.5%
Textile Industry	2	Light Manufacturing	Raw Coal/Natural Gas	1.5%
Timber Processing, Bamboo, Cane, Palm Fiber & Straw Products	2	Light Manufacturing	Raw Coal/Natural Gas/Gasoline	0.19%
Papermaking and Paper Products	2	Light Manufacturing	Raw Coal	1.69%
Other Manufacturing Industry	2	High Tech Industry	Raw Coal	0.15%
Production and Supply of Electric Power, Steam and Hot Water	2	Energy Production	Raw Coal	69.63%

Table 5. Feature analysis of the second cluster

lem of weakly correlated features; Elastic Net regression can combine the advantages of both Ridge regression and LASSO regression and achieve better compromise results.

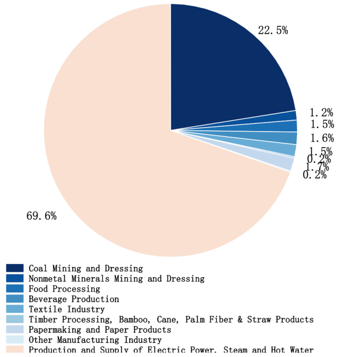

Figure 13. Distribution map of carbon emissions from the second cluster of energy sources.

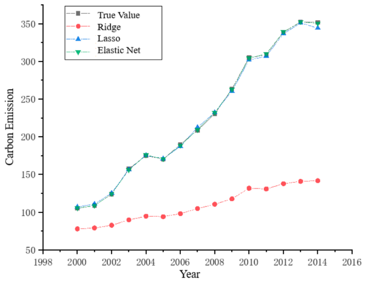

Figure 14. Fitting effect charts of three penalized regression methods.

In summary, Elastic Net regression performs best in terms of hyperparameter conditioning and sparsity, followed by LASSO regression. The Ridge regression method performs relatively poorly in terms of conditioning and sparsity, and the basic least squares method does not meet the requirements we need.

3.4.3. Prediction

The results of the Elastic Net regression were applied to forecast the 2015-2019 data, and the true and forecast values and the corresponding differences are summarized in Fig 15.

Year Data	2015	2016	2017	2018	2019
True	5.8063	5.7655	5.7647	5.6914	5.7531
Predict	5.8284	5.7107	5.7012	5.5943	5.6615
Difference	-0.0221	0.0548	0.0635	0.0972	0.0916

Figure 15. Elastic forecasts.

Mean Error: 0.0570. The mean error is close to zero and the predicted values tend to agree with the actual values overall.

Analysis of variance: 0.0023. The smaller variance indicates that the predicted values are relatively stable.

We compared the data from similar studies (Xie et al., 2024) (Shi, 2020)with the results of the current experiment and summarized comparison in Table 6.

Model	Area	MAPE(%)	RMSE	Max Error(%)
CNN-BFA	EU	4.6	7.8×10^{-2}	9.1
CNN-BFA	RGGI	5.3	8.4×10^{-2}	10.2
CNN-BFA	China	6.1	9.1×10^{-2}	11.3
LSTM-SVR	China	-	6.8×10^{-1}	-
DPR(Ours)	China	5.7	1.8×10^{-5}	2.38

Table 6. Prediction Comparsion with Baselines

Previous work used an LSTM-SVR hybrid deep learning model to analyze carbon emission data from different cities in China Economic Zone (Feda et al., 2024). Their dataset contained annual economic and social indicator data such as industrial added value, employment, and energy consumption from 2009 to 2018 for multiple cities. According to the data publicly available from that work, the LSTM-SVR hybrid model only had an RMSE of 0.682 on the test set. Other tested the forecasting performance of the optimized CNN-BFA hybrid deep learning model on three regional carbon emissions trading datasets, namely the datasets from the European Union Emissions Trading System, the Regional Greenhouse Gas Initiative, and the China Carbon Market (Feda et al., 2024). While the MSE of our proposed DBSCAN clustering coupled with penalized regression framework was 1.8309×10^{-5} , and R^2 was 0.9994. The results show that under the same experimental settings, DPR algorithm framework outperforms direct use of neural network modeling methods, and is better able to extract deep-level features of inter-industry relationships, thereby effectively reducing prediction errors and improving explanatory power.

In addition, studies also found that neural networks have difficulty effectively distinguishing highly homogeneous features in data (Feda et al., 2024). This also verifies the necessity of DPR framework using Density-based clustering to address the issue of multicollinearity.

4. Discussion

This study applied DBSCAN clustering algorithm to analyze the energy consumption data of 46 major industries in China from 2000 to 2019, and clearly extracted 16 characteristic classes. Among them, the second class clustered primarily around coal is an important source of carbon emissions. Targeted work can be carried out from the following aspects:

For the coal mining industry, clean coalbed methane extraction technology can be promoted to achieve simultaneous coal mining and coalbed gas extraction. The power system should accelerate the development of clean energy such as hydropower, wind power and solar power to gradually phase out small coal-fired power plants. It also needs to adopt smart grid technology to improve the proportion of renewable energy integration. The food industry should use highly efficient boilers and power generation equipment to realize industrial waste heat recycling and utilization.

The eighth class clustered primarily around gasoline includes the transportation industry. New energy vehicles and electric vehicles can be promoted. Delivery routes can be optimized to reduce trip frequency. Also, public transport network coverage needs to be expanded to reduce private car trips.

The steel industry, which consumes a large amount of coke, belongs to the ninth class clustered primarily around coke. Electric arc furnaces need to be widely adopted. Clean coking technology needs to be developed and slag-iron recycling is required.

Other measures for other industries include formulating clean industry standards to clarify emission reduction targets, promoting industrial waste heat utilization, implementing energy efficiency labeling certification, and building industrial parks to support technology research and development.

The above targeted measures respond to the needs of different classifications and industries, providing specific directions for worldwide low-carbon transition.

5. Conclusions

In summary, this study applied DPR framework based on DBSCAN clustering and penalized regression to analyze energy consumption and carbon emission data from 46 industries in China from 2000 to 2019. Important conclusions were drawn:

Through DBSCAN clustering to identify industry characteristics and penalized regression to identify major influencing factors, this framework deeply mines the intrinsic relationships between problems in an unsupervised learning and feature selection approach to provide quantitative basis for decision making.

This study validated the application value of DPR framework in identifying industry decarbonization opportunities and decision support. Future work will expand the sample scale and form a general analysis platform. At the same time, deep learning models can be combined to further improve it. Compared with traditional statistical models, this framework provides a new methodological idea. It provides new perspectives for widely applying data mining to urban optimization, resource allocation and other problems.

With the enhancement of data scale and computing power, large models are expected to further assist related fields in discovering more patterns and business value based on this framework. Specifically, the framework can be extended in the future in the following areas:

- Design modular algorithm architecture to support online learning and iterative optimization.
- Fully learn data hidden features using deep neural networks.
- Comprehensively mine complex relationships between multi-source and multi-dimensional data.
- Build a unified platform for multi-regional and multi-period analysis.
- Further promote the commercial application of intelligent decision support in broader areas.

Moving forward, this methodology holds promise to be scaled up and applied to decarbonization research in broader geographic regions. Its integrated approach leveraging machine learning and statistical modeling also has potential for addressing other multifaceted issues involving large, noisy datasets. With further refinement, the general framework could evaluate complex systems in areas like energy planning, environmental management and sustainable development. The ability to objectively define classifications, quantitatively assess relationships and control overfitting will benefit decision-making for transitioning towards low-carbon growth. Overall, this study contributes a data-driven technique for systematically analyzing multivariate problems with structural intricacies across diverse application domains. The code, data are available soon.

Impact Statement

Besides the value in carbon energy summarized in 4, I greatly appreciate in this paper is its demonstration of an important yet often overlooked point: sometimes, simpler models outperform more complex ones. In this study, a clustering-based approach followed by penalized regression proves to deliver superior performance compared to more intricate deep learning models, while providing more interpretable results. This aligns with findings in the scientific literature, where simpler approaches often excel in scenarios with limited data, high noise, or when interpretability is a priority. For instance, in bioinformatics, linear models such as elastic net regression have shown robust performance in predicting gene expression levels compared to deep neural networks when working with small datasets.

References

Feda, A. K., Adegboye, O. R., Agyekum, E. B., Hassan, A. S., and Kamel, S. Carbon emission prediction through the harmonization of extreme learning machine and INFO algorithm. *IEEE Access*, 12:60310–60328, 2024.

He, C., Duan, H., and Liu, Y. A neural network grey model based on dynamical system characteristics and its application in predicting carbon emissions and energy consumption in china. *Expert Syst. Appl.*, 266:126101, 2025.

Hoerl, A. E. and Kennard, R. W. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 42(1):80–86, 2000.

Li, G., Yuan, Y., Chen, X., Fu, D., and Jiang, M. Effectiveness of spatial measurement model based on SDM-STIRPAT in measuring carbon emissions from transportation facilities. *Energy Inform.*, 7(1):48, 2024.

Luo, J., Zhuo, W., Liu, S., and Xu, B. The optimization of carbon emission prediction in low carbon energy economy under big data. *IEEE Access*, 12:14690–14702, 2024.

Mozafari-Majd, E. and Koivunen, V. The adaptive τ-lasso: Its robustness and oracle properties. *CoRR*, abs/2304.09310, 2023.

Shi, C. Decoupling analysis and peak prediction of carbon emission based on decoupling theory. *Sustain. Comput. Informatics Syst.*, 28:100424, 2020.

Ullmann, T., Hennig, C., and Boulesteix, A. Validation of cluster analysis results on validation data: A systematic framework. *WIREs Data Mining Knowl. Discov.*, 12(3), 2022.

Xie, Q., Ballester-Berman, J. D., Lopez-Sanchez, J. M., Zhu, J., and Wang, C. Quantitative analysis of polarimetric model-based decomposition methods. *Remote. Sens.*, 8 (12):977, 2016.

Xie, T., Huang, Z., Tan, T., and Chen, Y. Forecasting china's agricultural carbon emissions: A comparative study based on deep learning models. *Ecol. Informatics*, 82:102661, 2024.

Zhang, Y., Lu, X., and Desmond, A. F. Variable selection in a log-linear birnbaum-saunders regression model for high-dimensional survival data via the elastic-net and stochastic EM. *Technometrics*, 58(3):383–392, 2016.

Zhu, H. and Li, D. A carbon emission adjustment model considering green finance factors in the context of carbon neutrality. *IEEE Access*, 12:88174–88188, 2024.

A. Appendix

A.1. DBSCAN Clustering

The DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm is a density-based clustering algorithm well-suited for clustering datasets with clusters of irregular shapes. The 46 industries exhibit complex clustering relationships in their energy usage patterns, which may form irregularly shaped clusters. DBSCAN can discover such clusters of any shape without specifying the number of clusters beforehand.

For a given dataset $Data = \{x_1, x_2, \ldots, x_n\}$, any two points x_i, x_j in the dataset are taken and the neighborhood parameters $(\epsilon, MinPts)$ are defined as follows:

Definition 1: ϵ is the radius of sample x_i, representing the extent of the circular region of the ϵ-neighborhood defined with x_i as the centre of the circle and ϵ as the radius of the domain.

Definition 2: $MinPts$ is the region density threshold of sample x_i, which determines whether x_i is used as a core point. When the sample value of the ϵ-domain range neighborhood of sample x_i is greater than the regional density threshold $MinPts$, it is determined that x_i is a core point.

In addition, DBSCAN uses several key concepts:

(1) **Core point:** For a point x_i, x_i is said to be a core object if the number of points contained within a certain radius r of it is greater than or equal to the parameter $MinPts$. Core objects are the key in the DBSCAN algorithm as they can form clusters.

(2) **Direct density access:** For two sample points x_i and x_j , x_i is called a directly density-reachable point of x_j if x_i is a point in the neighbourhood of x_j and x_j is a core object.

(3) **Density-reachable:** For two sample points x_i and x_j, if there exists a sequence of sample points $\{p_1, p_2, \ldots, p_n\}$, $p_1 = x_i, p_n = x_j$ and p_{i+1} is a direct density-reachable point of p_i, then x_j is said to be a density-reachable point of x_i.

(4) **Density-connected:** For two sample points x_i and x_j , x_i and x_j are said to be density connected if there exists a sample point x_k such that both x_i and x_j are density accessible points of x_k.

Based on the normalized data, DBSCAN clustering was performed combining the contour coefficient (SC) and sum of squared errors within clusters (SSE) for evaluation.

SC measures clustering goodness, with values closer to 1 indicating better results. **SSE** measures cluster internal similarity, with lower values signifying better results.

The contour coefficient (Ullmann et al., 2022) is that for the i-th sample point, let the cluster it belongs to be C_i then its average distance to all other sample points within cluster C_i is a_i, and its average distance to all other sample points in cluster $C_j (j \neq i)$ is b_i, then the contour coefficient s_i of this sample point can be defined as,

$$si = \frac{b_i - a_i}{\max(a_i, b_i)}$$

Where b_i denotes the average distance of the cluster with the smallest mean distance from this sample point in the set of clusters $C_j (j \neq i)$ to which this sample point belongs, i.e:

$$b_i = \frac{1}{|C_j|} \sum_{j \neq i, j \in C_j} d(x_i, x_j)$$

$d(x_i, x_j)$ denotes the distance between point x_i and point x_j. and a_i denotes the average distance of that sample point from other sample points within the same cluster:

$$a_i = \frac{1}{|C_i| - 1} \sum_{j \in C_i, j \neq i} d(x_i, x_j)$$

SSE is calculated as:

$$SSE = \sum_{i=1}^{n} \sum_{j=1}^{k} w_{ij} ||x_i - c_j||^2$$

Where n is the number of samples, k is the number of clusters, x_i is the i-th sample point, c_j is the centre of gravity of the j-th cluster, and w_{ij} is the weight of the i-th sample point belonging to the j-th cluster (1 means it belongs to that cluster, 0 means it does not).

A.2. Modeling of Regression

Unlike the STIRPAT model, our proposed method does not directly rely on its intrinsic assumptions, but rather proposes a more flexible analytical framework. DPR framework can not only identify the distinctive carbon emission profiles of different industries as demonstrated in this study, but also be extended to other domains such as air quality assessment and ecosystem modeling to systematically evaluate mitigation potentials at the subsystem level.

$$I = a_0 + \sum_{i=1}^{n} a_i x_i + e$$

Where a_0 is the constant term, e is the error term, x_i is the type of clustering, a_i refers to the regression coefficient and n is the number of clusters.

In order to establish an accurate linear equation to study the relationship between energy consumption and carbon emissions in 46 industries, we faced the challenge of multicollinearity, where multiple factors were interrelated and interfering with our analysis. To overcome this issue, we adopted a statistical modeling technique called penalized regression. This technique introduces penalty terms into the linear regression model to control model complexity and enhance its generalization performance.

In our study, we employed several common penalized regression algorithms, including Ridge Regression, Lasso Regression, and Elastic Net Regression, modeling as three parameters in the training process. Specifically:

Ridge regression employs L2 regularization to address potential multicollinearity in industry energy data. It shrinks parameter estimates to improve model stability.

- **Loss function:** $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
- **Object function:** $\min(\sum_{i=1}^{n}(y_i - \overline{y}_i)^2 + \lambda \sum_{j=1}^{p} \beta_j^2)$

Lasso regression uses the L1 regularization term (Mozafari-Majd & Koivunen, 2023), which is the sum of the absolute values of the parameters multiplied by a penalty factor. l1 regularization is sparse, i.e. it allows feature selection by estimating some parameters as zero. lasso regression is useful in situations with a large number of features and where feature selection is required.

- **Loss function:** $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
- **Object function:** $\min(\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} |\beta_j|)$

Elastic Net Regression (ENR) is a linear regression method that combines ridge regression and lasso regression Section (Zhang et al., 2016). In Elastic Net Regression, the objective function consists of two components: the loss function and the penalty term. The penalty term has two components: L1 regularization and L2 regularization. L1 regularization limits the complexity of the model by the sum of the absolute values of the parameters and achieves the effect of feature selection. L2 regularization limits the growth of the parameters by the sum of the squares of the parameters and reduces the variance of the parameter estimates.

- **Loss function:** $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
- **Object function:**

$$\min(\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j|) + \lambda_2 \sum_{j=1}^{p} \beta_j^2$$

where y_i is the actual value of the observation, $\hat{y}_i$ is the predicted value of the model, β_j is the regression coefficient, p is the number of independent variables, and $\lambda, \lambda_1, \lambda_2$ are hyperparameters that control the degree of regularisation.

We applied these penalized regression models to industrial data to identify key influencing factors and address the challenges associated with high-dimensional data. To assess the effectiveness of these models, we introduced two evaluation metrics:

MSE measures the average squared difference between the actual observed values and the predicted values produced by the regression model.

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

R^2 quantifies the proportion of the variance in the dependent variable (e.g., carbon emissions) that is explained by the independent variables (e.g., energy consumption and other factors) in the regression model.

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

Where n is the number of samples, y_i is the actual value of the i-th sample, $\hat{y}_i$ is the predicted value of the i-th sample, and $\overline{y}$ is the average of the actual values.

Sparsity s refers to the proportion of non-zero elements in the parameter or feature vector, i.e. how much useful information is retained. The addition of a regularization term can make the parameter or feature vector more likely to have zero elements and thus achieve sparsity.

$$s = \frac{\sum_{i=1}^{n} I(\beta_i \neq 0)}{p}$$

Let s denote sparsity, where $s \in [0, 1]$ represents the proportion of non-zero elements in a parameter/feature vector. n denotes the number of samples, p denotes the number of parameters/features. I is the indicator function that takes value 1 if a coefficient is non-zero, and 0 otherwise.A perfectly dense vector has s = 1, while a perfectly sparse vector has s = 0.

Socio-economic Sectors	Classification	Industry Category	Major Energy Distribution
Farming, Forestry, Animal Husbandry, Fishery and Water Conservancy, Other Minerals Mining and Dressing Construction	1	Primary Industry Energy Production Construction	Diesel/Gasoline
Wholesale, Retail Trade and Catering Services	1	Service Industry	Diesel/Gasoline
Coal Mining and Dressing, Nonmetal Minerals Mining and Dressing	2	Energy Production	Raw coal/Natural Gas
Food Processing, Beverage Production, Textile Industry, Timber Processing, Bamboo, Cane, Palm Fiber & Straw Products, Papermaking and Paper Products	2	Light Manufacturing	Raw coal/Natural Gas
Other Manufacturing Industry, Production and Supply of Electric Power, Steam and Hot Water	2	New Energy Production	Raw coal/Natural Gas
Petroleum and Natural Gas Extraction	3	Energy Production	Refinery Gas/Natural Gas
Ferrous Metals Mining and Dressing	4	Energy Production	Other Gas/Coke
Nonferrous Metals Mining and Dressing	5	Energy Production	Other Gas/Gasoline
Food Production, Leather, Furs, Down and Related Products, Furniture Manufacturing	6	Light Manufacturing	Natural Gas
Chemical Fiber Metal Products, Equipment for Special Purposes, Electric Equipment and Machinery, High-tech Industry, Production and Supply of Gas	6	Heavy& High Tech Manufacturing	Natural Gas
Tobacco Processing	7	Light Manufacturing	Other Coal Washing
Garments and Other Fiber Products, Printing and Record Medium Reproduction	8	Light Manufacturing	Gasoline/Coal/Natural Gas
Rubber Products, Plastic Products, Production and Supply of Tap Water	8	Heavy Manufacturing	Gasoline/Coal/Natural Gas
Electronic and Telecommunications Equipment, Instruments, Meters, Cultural and Office Machinery	8	Hign Tech Industry	Gasoline/Coal/Natural Gas
Urban, Rural	8	Household	Gasoline/Coal/Natural Gas
Cultural, Educational and Sports Articles	9	Light Manufacturing	Coke
Smelting and Pressing of Ferrous Metals Coke, Ordinary Machinery	9	Heavy Manufacturing	Coke
Scrap and waste	9	High-tech Industry	Coke
Petroleum Processing and Coking	10	Energy Production	Clean Coal/Other Washed Coal
Raw Chemical Materials and Chemical Products	11	Heavy Manufacturing	Natural Gas
Medical and Pharmaceutical Products	12	Light Manufacturing	Other Coal Washings/Raw Coal/Gasoline
Nonmetal Mineral Products	13	Heavy Manufacturing	Process
Smelting and Pressing of Nonferrous Metals	14	Heavy Manufacturing	Natural Gas/Coke/Other Gases
Transportation Equipment	15	Heavy Manufacturing	Natural Gas/Other Coal Washing/Gasoline
Transportation, Storage, Post and Telecommunication Services	16	Service Industry	Kerosene/Diesel/Gasoline

Table 7. Clustering result