ISSN 2616-5775 Vol. 8, Issue 6: 84-89, DOI: 10.25236/AJCIS.2025.080610

Study on the Influence of Process Parameters in Cold-Rolled Strip Using Statistical and Machine Learning Methods

Jiao Tang[1], Yixuan Zhang[1], Yanzhuo Wu[1], Qingkun Yu[2,*]

[1]*School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan, China, 114051*
[2]*College of Science, University of Science and Technology Liaoning, Anshan, China, 114051*
[*]*Corresponding author: yuqingkun@163.com*

Abstract: *In the steel industry, the mechanical properties of cold-rolled strips are crucial for product quality and cost. This study aims to optimize the continuous annealing process parameters. Data preprocessing involved identifying and removing outliers using box-and-line plots and addressing missing values with a decision tree-filling method. Descriptive analysis was performed to determine the statistical characteristics of the process parameters, and normality was assessed using P-P plots, Q-Q plots, and histograms, revealing that the data did not follow a normal distribution. The Spearman correlation coefficient was used to examine parameter correlations, identifying those that significantly affect strip hardness based on their significance levels. A Decision Tree machine learning model was then employed to analyze feature importance, identifying six key parameters influencing strip hardness. The top three factors were the temperature of the fast-cooling furnace (29.1%), carbon content (25.1%), and the temperature of the quenching furnace. Notably, the importance of fast-cooling furnace temperature and carbon content was equal, while strip speed was the least influential factor, accounting for only 3.2%. This research provides essential insights and practical references for optimizing the production process and quality control of cold rolled strip steel.*

Keywords: *Process Parameter Analysis, Statistical Characterization, Strip Hardness, Decision Tree Modeling*

1. Introduction

In the contemporary steel sector, cold-rolled steel strips are of crucial importance as their mechanical properties are deeply intertwined with the quality and lifespan of downstream products, thereby playing a decisive role in the market competitiveness of end products. Continuous annealing, an essential thermal treatment in the production of cold-rolled strips, is key to optimizing the hardness, toughness, and other performance indices of steel.

This research is highly significant as it endeavors to accurately identify the key process parameters that govern the performance of cold-rolled strips. By doing so, it provides steel manufacturers with the means to enhance the stability of product quality. Downstream industries such as automotive and construction can obtain more reliable materials. Additionally, optimizing these parameters leads to remarkable energy savings and waste reduction, promoting sustainable manufacturing practices and minimizing the environmental footprint. The insights derived from this study also have the potential to fuel innovation in steel production technologies, facilitating the development of more advanced annealing processes and materials and thus helping the steel industry to maintain a competitive edge in the international market.

To achieve these goals, the study focuses on comprehensively analyzing the multi-dimensional data from the continuous annealing process. State-of-the-art statistical and machine learning techniques are employed. The novelty of this research lies in the seamless integration of diverse analytical methods. Box-and-line plots are used for outlier detection and removal to ensure data integrity. A clear understanding of data distribution is obtained through comprehensive statistical characterizations, which simplifies the identification of anomalies and lays a solid foundation for further analysis. Histograms, Q-Q plots, and P-P plots are utilized for normality assessment, which is essential for evaluating the applicability of linear correlation methods. The incorporation of decision tree algorithms is particularly

ISSN 2616-5775 Vol. 8, Issue 6: 84-89, DOI: 10.25236/AJCIS.2025.080610

innovative as they can effectively handle the production environment's complexity by considering the nonlinear interactions among parameters. These algorithms can precisely identify the key parameters influencing hardness through feature importance analysis, presenting a pioneering approach to optimizing cold-rolled strip production.

2. Data Sources and Preprocessing

This study centers on the effect of continuous annealing process parameters on the properties of cold-rolled strip steel. The data used was sourced from the website https://cxcy.upln.cn. This source details 12 key performance indicators in cold strip production, including strip thickness, strip width, carbon content, silicon content, strip speed, furnace temperature, warming and cooling furnace temperature, aging furnace temperature, fast cooling furnace temperature, quenching furnace temperature, and leveling machine tension.

After obtaining the original data, the strip specification data is easily disturbed by various complex factors, such as fluctuations in the accuracy of the measuring equipment and negligence in manual recording; this leads to the data appearing in abnormal values or inconsistent before and after the record. Outliers can impact the accuracy of subsequent analyses, and data cleaning will be performed in this paper to rule out this risk.

First, outliers are detected using the interquartile range box line [1] (IQR) and processed by nulling. Once an outlier is detected, it is instantly zeroed and processed to purify the dataset and remove “noise.” Figure 1 shows the box plots of the 12 process parameters for detecting outliers. Box and line diagram.

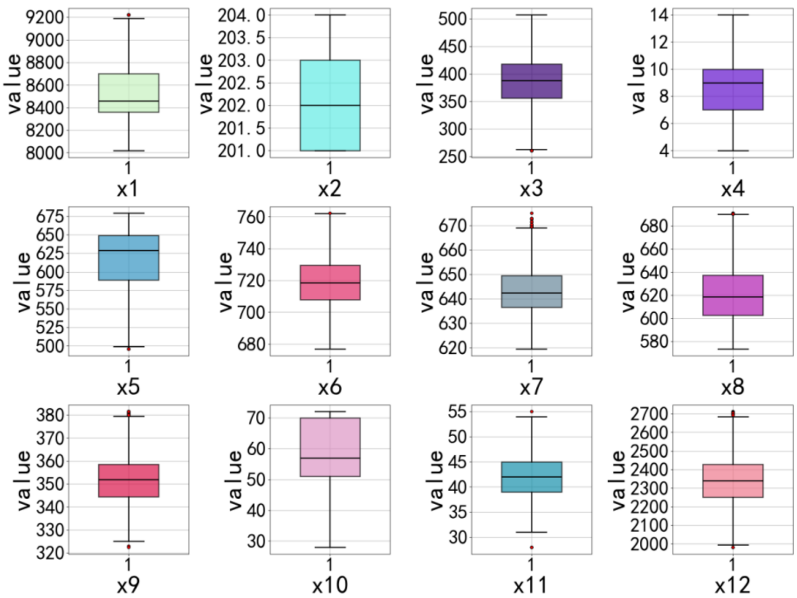

Figure 1. Box and line diagram

Where x1 (strip thickness), x2 (strip width), x3 (carbon content), x4 (silicon content), x5 (strip speed), x6 (heating furnace temperature), x7 (homogenizing furnace temperature), x8 (warming and cooling furnace temperature), x9 (over-aging furnace temperature), x10 (fast-cooling furnace temperature), x11 (quenching furnace temperature), x12 (flattening machine tension), y (hardness).

This operation improves the data quality and builds the foundation for introducing decision trees to fill in missing values. Use the decision tree filling module of the existing analysis software to fill in the null values. In addition, using statistical descriptive methods provides a comprehensive insight into the distribution of the data and the range of values. The following is Table 1.

In the results of descriptive statistics for each variable, the data showed diverse characteristics after the decision tree was filled. It is worth noting that the data of strip thickness and leveler tension fluctuate a lot, and their variances are 70099.66 and 21430.11. Meanwhile, from the data pattern, the kurtosis of strip thickness is -0.468, which is relatively small, indicating that its data distribution is relatively smooth, and there is no apparent phenomenon of sharp peaks or thick tails; its skewness is 0.477, which shows right skewness, implying that there is a longer trailing tail on the right side of the data, and the right side

ISSN 2616-5775 Vol. 8, Issue 6: 84-89, DOI: 10.25236/AJCIS.2025.080610

of the data is relatively dispersed. In addition, the coefficient of variation of silicon content is 0.235, which is relatively large, indicating that the degree of dispersion of the data of this variable is more prominent among the variables, the stability of the data is poorer, and the difference of the values of silicon content between different samples is relatively more apparent.

Table 1. Descriptive statistics results after filling the decision tree

variable name	sample size	maximum values	minimum value	average value	standard deviation	upper quartile	variances	kurtosis	skewness	coefficient of variation
x1	1000	9220	8020	8542.31	264.763	8460	70099.664	-0.468	0.477	0.031
x2	1000	204	201	202.105	0.88	202	0.775	-0.9	0.227	0.004
x3	1000	507	261	383.683	48.251	388	2328.137	0.009	-0.393	0.126
x4	1000	14	4	8.736	2.055	9	4.223	-0.044	0.336	0.235
x5	1000	679	496	614.589	42.152	629	1776.805	-0.092	-0.855	0.069
x6	1000	762.2	677	719.19	16.923	718.6	286.4	-0.025	0.126	0.024
x7	1000	675	619.5	643.559	9.502	642.5	90.281	0.342	0.707	0.015
x8	1000	691	573.5	622.49	26.428	618.5	698.438	-0.105	0.67	0.042
x9	1000	381.5	322.5	352.75	13.099	352	171.594	-0.302	0.411	0.037
x10	1000	72	28	59.783	9.604	57	92.23	-1.239	-0.153	0.161
x11	1000	55	28	41.813	4.513	42	20.364	-0.356	-0.058	0.108
x12	1000	2711.793	1982.077	2348.85	146.39	2339.191	21430.107	-0.246	0.174	0.062
y	1000	650	540	596.81	20.143	600	405.73	-0.066	-0.798	0.034

3. Normality Testing and Correlation Assessment of Process Data

Normality tests (e.g., Q-Q plots, P-P plots, and histograms) were performed to determine if statistical distributions could be analyzed using linear relationships between parameters [2]. Figures 2 through 4 below demonstrate the results of the normality test.

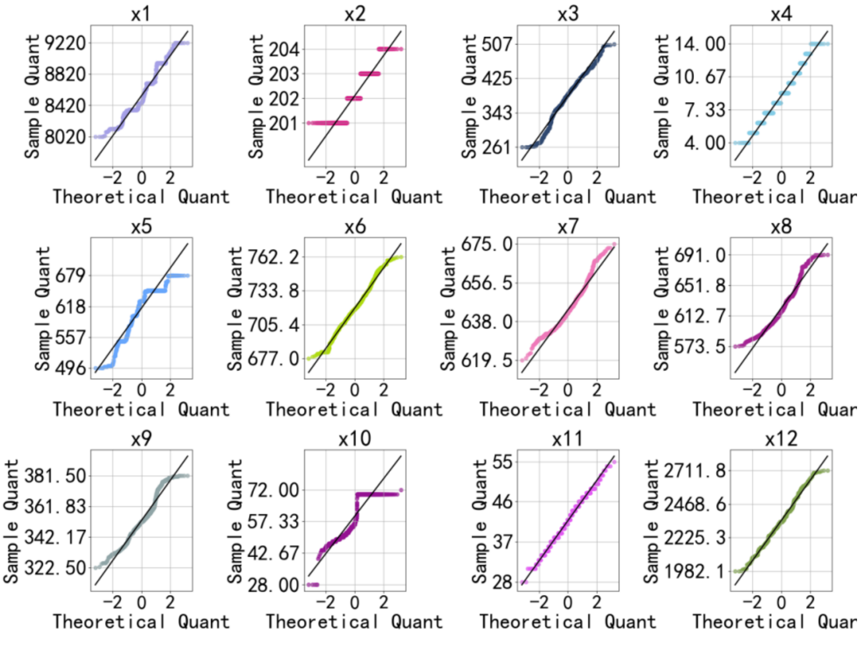

Figure 2. Q-Q normality test

In Figure 2, the data points of most variables are close to the straight line but with deviations, such as strip thickness deviating significantly at both ends and strip speed deviating in the S-shape, showing that the data do not fully conform to the theoretical distribution.

In Figure 3, the data points of each variable are off from the diagonal line, like the data points of the fast-cooling furnace temperature, which are mostly above the diagonal line. This indicates a discrepancy between the actual and theoretical distributions.

ISSN 2616-5775 Vol. 8, Issue 6: 84-89, DOI: 10.25236/AJCIS.2025.080610

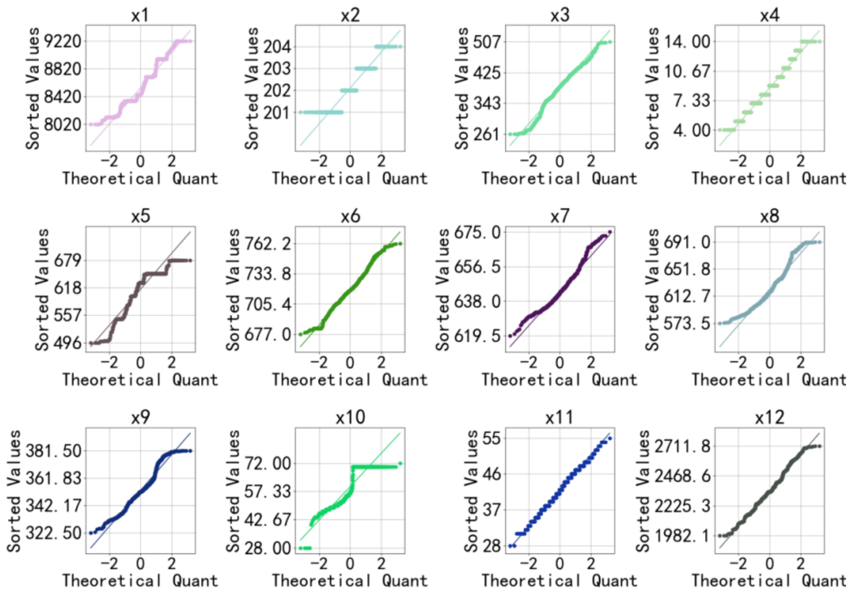

Figure 3. P-P normality test

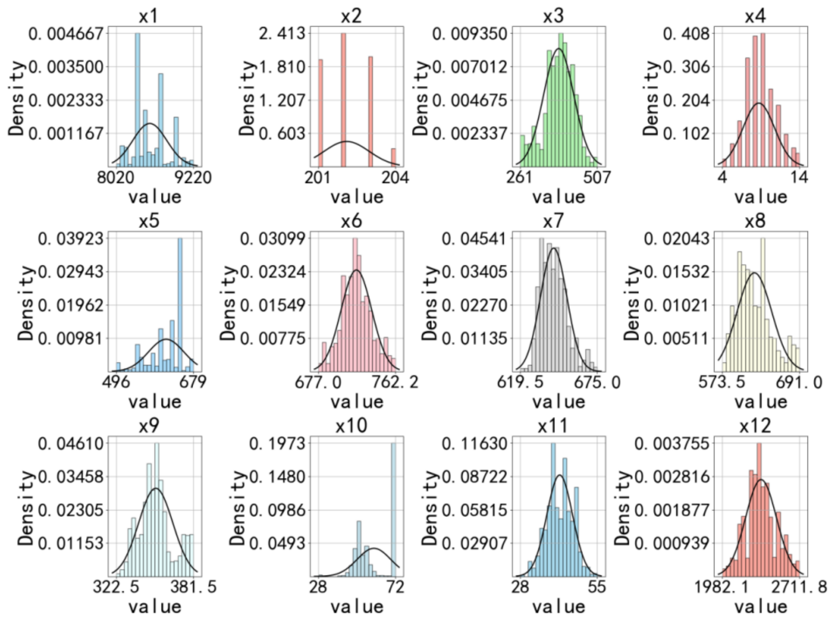

Figure 4. Histogram normality test

In the histogram of Figure 4, the variables are distributed in various patterns; the carbon content is right-skewed and highly dispersed, and the over-aging furnace temperature is left-skewed and more concentrated, reflecting the complexity and diversity of data distribution.

An in-depth analysis of the P-P, Q-Q, and histograms shows that the data do not conform to the characteristics of a normal distribution, which is also corroborated by the histograms of the distribution of the indicators in Figure 4. Further, considering that Pearson's correlation coefficient [3] is more suitable for normally distributed datasets, Spearman's correlation coefficient was used to assess the correlation between the independent and dependent variables. The following is Figure 5.

ISSN 2616-5775 Vol. 8, Issue 6: 84-89, DOI: 10.25236/AJCIS.2025.080610

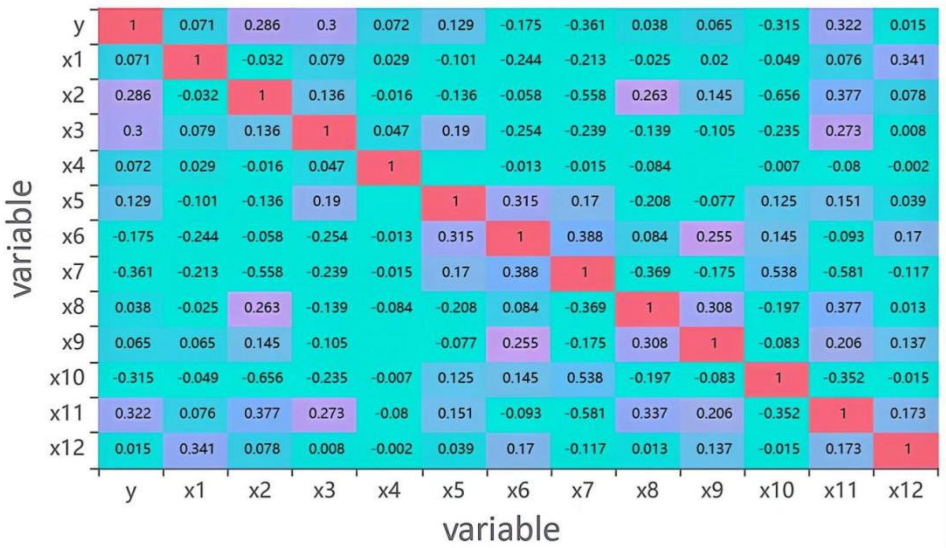

Figure 5. Spearman correlation coefficient matrix

Figure 5 shows that the effect of variables carbon content, strip speed, heating furnace temperature, uniform heating furnace temperature, fast cooling furnace temperature, and quenching furnace temperature on the hardness is highly significant. The correlation between hardness and the seven variables is strong: strip width, carbon content, strip speed, heating furnace temperature, uniform heating furnace temperature, fast cooling furnace temperature, and quenching.

4. Analysis of Hardness Influencing Factors

This paper adopts the decision tree regression algorithm to identify the key process parameters that most impact hardness. The algorithm was used to analyze the impact of multiple process parameters on hardness. The decision tree algorithm identifies seven key process parameters that most impact hardness. The proportion of importance of each feature is shown in Table 2.

From Table 2, it can be seen that carbon content and fast cooling furnace temperature have the highest characteristic importance [4] of 25.10% and 29.10%, respectively, which indicates that these two features contribute the most to the prediction of hardness of the target variable. Other important characteristics include warm and cool furnace temperature (12.90%) and quench furnace temperature (8.70%). Strip width does not contribute to the prediction, and its significance is 0%. The strip widths were therefore rounded off from seven and are also given in Table 3.

Table.2. Characteristic importance ratios

Feature name	Characteristic importance
x1	5.70%
x2	0.00%
x3	25.10%
x4	1.50%
x5	3.20%
x6	4.30%
x7	4.90%
x8	12.90%
x9	2.70%
x10	29.10%
x11	8.70%
x12	1.90%

Table 3. Results of model evaluation

	MSE	RMSE	MAE	MAPE	R^2
training set	150.695	12.276	8.922	1.504	0.628

Using the decision tree [5] algorithm to analyze the data in Tables 2 and 3, the six key process parameters with the most significant impact on hardness were successfully identified. These parameters,

ISSN 2616-5775 Vol. 8, Issue 6: 84-89, DOI: 10.25236/AJCIS.2025.080610

in descending order of importance, are the fast-cooling furnace temperature (29.1%), carbon content (25.1%), quenching furnace temperature (8.7%), homogeneous heating furnace temperature (4.9%), heating furnace temperature (4.3%), and strip speed (3.2%). The corresponding model evaluation indexes, such as MSE of 150.695 and RMSE of 12.276, demonstrate that the Decision Tree Machine Learning Model has good performance and explanatory power. It can accurately identify the six process parameters that play a key role in the hardness of the strip and thus provide key guidelines for optimizing the cold-rolled strip steel production process.

5. Conclusions

This study focuses on the effect of continuous annealing process parameters on the hardness of cold rolled strip steel, using a combination of statistical analysis and machine learning to explore in depth. A large number of process parameters are recognized one by one with the help of a box diagram, and outliers are manually blanked out. The decision tree filling method is employed, involving multiple trials of adjusting the decision tree parameters to handle missing values. Diverse statistical indicators are calculated for numerous process parameters when performing descriptive analysis. Normality was tested using P-P plots, Q-Q plots, and histograms. Finally, Spearman's correlation coefficient is used to analyze the parameters that significantly affect strip hardness by constructing a complex matrix of correlation coefficients containing several process parameters.

The results of this research are not only highly feasible to be applied in the field of steel manufacturing but also of great practical significance. Iron and steel companies can optimize their processes with the parameters obtained at the end of the process and improve their efficiency through precise temperature control, strict carbon control, upgrading of temperature control equipment, and optimization of speed. In the future, the range of parameters can be expanded to include factors such as fluctuations in the composition of raw materials and the degree of wear and tear of equipment, and complex models can be used to dig deeper into the relationship and improve the accuracy of forecasts.

References

[1] Fu Wenlong, Zhao Yibo, Fu Yuchen, et al. Prediction of dissolved gas volume fraction in transformer oil based on improved box and line diagram and ISMAHHO-KELM [J/OL]. High Voltage Electrical Apparatus, 2024, 1-13 [2024-11-30].

[2] Zong Xuping, Yao Yulan. Quickly check the statistical distribution of data using Q-Q and P-P plots [J]. Statistics and Decision Making, 2010, (20): 151-152.

[3] Ejegwa, Paul Augustine, Kausar, Nasreen, Aydin, Nezir, et al.A new Fermatean fuzzy Spearman-like correlation coefficient and its application in evaluating insecurity problem via multi-criteria decision-making approach [J]. Heliyon, 2024, 10(22).

[4] Zidong Xu, Zhenyong Zhang. Power system anomaly detection method based on feature importance of unbalanced high-dimensional data [J]. Control Engineering, 2024, 31(11): 2029-2035.

[5] Yi Jingjing, Li Xingting, Pu Chunrong, et al. Risk warning model for postoperative adverse pregnancy outcomes in patients with cervical insufficiency based on decision tree algorithm [J/OL]. Chongqing Medicine, 2024, 1-8 [2024-11-30].