Industrial Engineering and Applications
L.-C. Tang (Ed.)

doi:10.3233/ATDE230057

A Two-Step Method of Statistical Difference Testing and K-Means Clustering for Identifying Quality Factors of Small Steel Bars

Dong-Hee Lee[a,1] and Kwang-Ho Jeong[a]
[a] *Department of Industrial Engineering, Sungkyunkwan University*
ORCiD ID: Dong-Hee Lee https://orcid.org/0000-0001-8549-8992, Kwang-Ho Jeong https://orcid.org/0009-0001-5694-9667

Abstract. As profit of steel enterprises is getting smaller and comparison is getting harder, in steel manufacturing industry quality inspection and control have become main measures for getting over the difficulties. In sequential manufacturing process, failure occurred in preceding process make final quality worse. Detecting the factor which affects to final quality is necessary to make process stable and efficient. Existing studies to improve the surface quality of small steel bars assume that these quality factor have been identified and focus on diagnosis of defects. However, there are no attempts to validate the quality factor based on data. In this paper, we attempt to verify the quality factor based on data using statistical and data-mining techniques. To get over practical problems from variation of quality due to operation date and merge measurement of quality, we suggest method using statistical significance difference and using k-means clustering(k=2). Method using statistical significance difference considers the overall tendency about quality and method using k-means clustering consider the tendency toward outlier. Quality factors of small steel bar can be detected by using both methods serially. And we apply this method to real world data in case study.

Keywords. Quality factor, Steel rolling, Small steel bar, Data-mining, Statistic analysis, Merged measurement

1. Introduction

With the serious excess of iron and steel production in the world, the profit of steel enterprises is getting decreased, and the competition is getting increased[1]. In the steel-manufacturing industry, quality inspection and control have been main measures for getting over the difficulties. Because a steel defect is deemed to be one of the main causes of the production cost increase, controlling the quality of steel products is inevitable[2].

Steel-manufacturing process is mainly composed of 4 sequential processes: iron making, steel making, continuous casting and rolling process. If a failure occurs in preceding process, the quality of final products may be threatened and be greatly affected.

[1] Dong-Hee Lee, Corresponding author, Department of Industrial Engineering, Sungkyunkwan University, 2066 Seobu-ro, Jangan-gu, Suwon 16419, Republic of Korea, E-mail: dhee@skku.edu

The condition of the assets and stability of the processes are key factors in manufacturing systems, which determine quality of the product and assure continuity of the production[3]. And efficient process performance and stable quality are necessary[4]. In this research, we define the factor which affects final product quality as 'Quality factor'. Identifying quality factors is needed to make process stable and efficient.

Existing studies to improve the surface quality of small steel bars assume that these quality factor have been identified and focus on diagnosis of defects[5], [6]. However, there are no attempts to validate the quality factor based on data. In this paper, we attempt to identify quality factors that have significant effect on a reprocessing ratio, which is a one of critical quality index in steel manufacturing process. The quality factors are identified by collecting a large amount of operational data from the steel manufacturing process and using statistical and data-mining techniques. Each method has macro and micro perspectives to identify quality factors.

The remainder of the paper is organized as follows. In Section II, Practical difficulties in identifying the quality factors are introduced. In Section III, the proposed method for identifying the quality factors and its case-study are explained in Sections III and IV, respectively. Finally, conclusion is given in Section V.

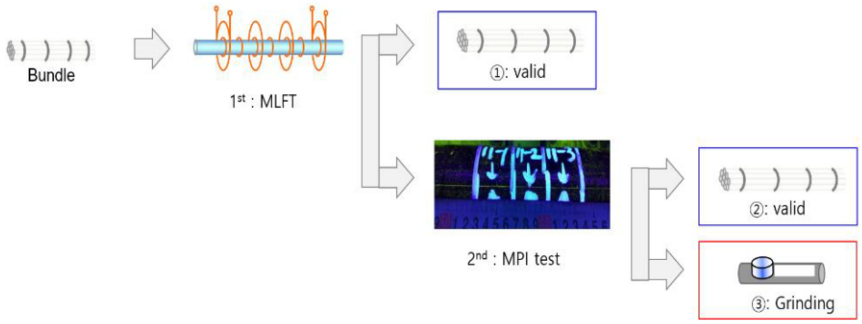

Figure 1. Quality inspection process in small bar manufacturing

2. Problem Statement

First practical problem we match is that reprocessing ratio, quality index, is measured by merged. **Figure 1.** shows flow chart of Quality inspection process in small steel bar manufacturing. The quality inspection stage of small steel bar consists of two stages. The first stage is MLFT(Magnetic Leakage Flus Testing). It is a non-destructive inspection method that detects the leakage magnetic flux of the defective part using a magnetism. Products that have been judged to be defective in the first stage will move to the second stage. The second stage is MPI(Magnetic Particle Inspection). It is used for the detection of surface and near-surface flaws in ferromagnetic materials and is primarily used for crack detection. Products that have been judged to be defective even in the second inspection are reprocessed to grind the surface. In our case example, several small steel bars are merged into a bundle and reprocessing ratio is measured for each bundle. In **Figure 1.**, ① is the number of validated small steel bars in 1st inspection. ② is the number

of validated small steel bars in 2nd inspection. And ③ is the number of small steel bars that are grinded. Reprocessing ratio follows Eqs. (1):

$$reprocessing\ ratio\ value = \frac{③}{① + ② + ③} \tag{1}$$

The bundle, which is a group of several small steel bars, takes reprocessing ratio after quality inspection process. Thus, Small steel bars that belong to the same bundle have same reprocessing ratio. We call this problem as "a merged measurement problem".

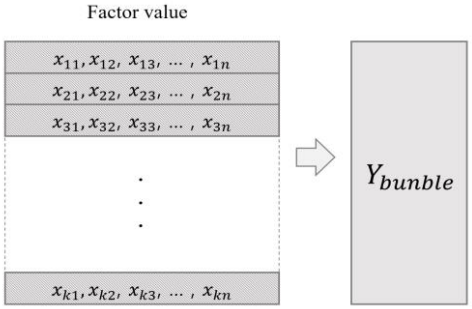

Figure 2. Relationship between factor value and reprocessing ratio

This merged measurement problem imposes many restrictions on the use of data-mining techniques to derive quality factors of small steel bar. Supervised learning is commonly used to find quality factors. Each record (or observation) should have both values for a factor and a response variable. However, supervised learning is not available in our case example due to the merged measurement problem because each small steel bar has only factor values without reprocessing ratio. unsupervised learning can be used for anomaly detection. However, it has a loss of cost because it does not use the already measured reprocessing ratio data. **Figure 2.** explains relationship between factor value as input x and reprocessing rate as output y. In **Figure 2.**, each row is data that each product has. x_{ij} means j th factor value of i th small steel bar.

The second practical problem is the effect of production timing on products in the steel process. Depending on the operation date, changes in external environments such as external temperature and humidity and facility alignment occur. Since steel manufacturing facilities are open facilities that are not isolated from the external environment, the external environment may affect the condition of the production line. In addition, since it is a large facility, it is difficult to control the change in facility alignment for every operation period. Thus, there is a problem that the timing of production must be considered when deriving quality influencing factors. We call this problem as "variation due to the operation date".

Therefore, to identify quality factors of small steel bars, it is necessary to consider both "the merged measurement problem" and "the variation due to the operation date".

3. Methodology

In this paper, we present a two-step method of statistical analysis and data-mining to identify quality factors for small steel bars in considering both "the merged measurement problem" and "the variation due to the operation date".

3.1. Using Statistical Significance Difference for Quality Factor Detection

The statistically significant difference testing compares whether the statistical difference in factor values between groups with high reprocessing ratio and the group with low reprocessing ratio is statistically significant. And the fact that the factor values are in the same group implies the factor values have same reprocessing ratio. To verify the statistical significance of the mean difference in factor values between the two groups, t-test for the numerical variable and chi-square test for the categorical variable are adopted. If the mean difference in factor values between the two groups, divided according to the reprocessing ratio, is statistically significant, the factor is assumed to be a quality factor.

In order to consider the remaining practical problem, the statistical difference testing for each operation date is conducted by forming groups with high and low reprocessing ratio for each operation date. If a comparative group is formed by production period, the difference between the two groups can be considered considering the effect of changes in the external environment and the production period of facility characteristics. If testing groups are divided by operation date, the difference of factor values between the two groups can be compared considering the effect of changes in the external environment and the operation date of facility alignments. As a result of analyzing all operation date, factors that repeatedly show statistically significant differences in multiple operation date are derived as quality factors.

There are problems to be considered in statistical difference testing. First, the level of product quality management is not controlled constantly by operation date. Thus, it is necessary to consider the difference between high reprocessing ratio and low reprocessing ratio by operation date. In the operation date when the product quality management level is high, there will be little difference in reprocessing ratio between high and low groups. The difference in factor values between groups with little difference in reprocessing ratio cannot be assumed to the cause of the difference in reprocessing ratio. Therefore, we conduct the statistical difference testing only when the difference is over a certain level. Second, it is the number of data by group. In the significant difference testing that verifies the statistical difference in factor values between groups, the number of samples often affects the results. Therefore, to obtain a significant p-value in the statistical difference testing, the testing must be conducted on the groups with a certain number of data or more. That is, the testing should be conducted except for group with the same reprocessing ratio which is less than certain number of data.

Through statistical difference testing, it is possible to first screen quality factors macroscopically by grasping the tendency of reprocessing ratio for the overall factor values. However, since this method is vulnerable to the outlier, microscopic analysis is needed to consider the outlier. Therefore, we introduce a microscopic analysis method for detecting the outlier of factor values using K-means clustering(K=2).

3.2. Using K-means Clustering(K=2) for Quality Factor Detection

Using K-means clustering(K=2) is a method of calculating the composition ratio of factor values by factor using K-means clustering and comparing it with the reprocessing ratio. Since the factors are clustered within the same reprocessing ratio, the group has identical reprocessing ratio. That each data in group has the identical reprocessing ratio implies that each data in group also has the identical operation date. So, it is an analysis method that considers both "the merged measurement problem" and "the variation due to the operation date".

The method using K-means clustering is based on the following assumptions. We do not know whether the individual defect of small steel bars exist, but the defect ratio is provided through the reprocessing ratio. If a quality factor has an outlier, defect will occur, and if the ratio of the outlier to the factor is analogous to the reprocessing ratio, the factor can be assumed as the cause of the quality problem. If such an event occurs repeatedly at various reprocessing ratio, the factor is derived as a quality impact factor. Dividing the factor value into two groups using K-means clustering(K=2) has two effects: 1) K-means clustering will separate the outliers and the values of the normal distribution into different groups in response to the outlier. 2) If there is no outlier in the distribution of the factor values, the two groups will have a similar ratio. Therefore, in the situation where defect rate is constantly managed low, it is possible to classify in response to the outlier of the factor value using K-means clustering.

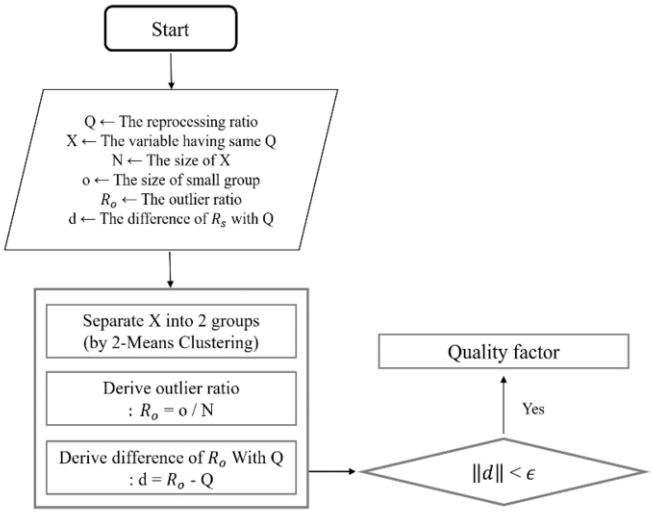

Figure 3. A flowchart of the using k-means clustering for quality factor detection

Figure 3. is a flowchart of the using K-means clustering for quality factor detection. It is repeatedly applied for each of the factor values of all bundles. The factor values are divided into two groups using K-means clustering. And the ratio of the number of small groups between two divided by the number of total factor values is defined as the outlier ratio. Finally, if the difference between the outlier ratio and the reprocessing ratio is within a preset acceptable range, the factor is assumed to be a quality factor. This process

is repeated for all reprocessing ratio. And factors that are repeatedly derived in multiple reprocessing ratio are derived as quality factors.

Also, there is problem to consider when using K-means clustering method. When the factor value is constantly increases or decreases over time, applying the method is impossible. The situation where these factors have an outlier exist only when there is a problem with given data. Because the two groups must have similar sizes.

3.3. Serial Application of Both Means

We propose a method of using the statistical difference testing and using K-means clustering in series. First, a statistical difference testing is conducted to identify the tendency of the product quality to the overall factor value from a macroscopic point of view, and the factor with the tendency to the product quality is screened. Continuously, through method using K-means clustering to detect factors that affect product quality from a microscopic perspective, quality factors are selected.

4. Illustrative Example

4.1. Description of Dataset

The raw dataset was collected from small steel bar manufacturing process in Hyundai-Steel co. from Jan 2021 to June 2022. The dataset contains 11 620 samples and 407 features. Each sample corresponds to each small steel bar. 11 620 small steel bars are divided into 347 bundles by 347 reprocessing ratio values. 347 groups are divided into 36 operation date. And, 27 features that increase or decrease consistently over time, 61 features that are derivative from other features such as increasing rate between two features, and 27 features with same value in each column were removed, and the remaining 292 features were used for analysis. The 292 features consist of 269 numerical features and 23 categorical features.

4.2. Experimental Settings

The two methods proposed to derive quality factors require each hyperparameter. The hyperparameters of the testing are the under-bound number of data, the reprocessing ratio difference between highest and lowest ratio groups, and the p-value criterion. The under-bound was set to 20. With the help of engineers, the reprocessing ratio difference between groups was set to 3, and the p-value criterion was set to 0.05. The hyperparameter of the method using K-means clustering is an acceptable range for the difference between the reprocessing ratio and the outlier ratio. The acceptable range was applied differently depending on the size of the reprocessing ratio. In statistical difference testing, Screening was performed on factors detected below 30% of the total operation date. In method using K-means clustering, factors detected more than 10 times were suggested as quality factors.

4.3. Result

Table 1. Detection times of feature for both methods

Feature	Method		Feature	Method	
	1)statistical difference testing	2)Using K-means clustering		1)statistical difference testing	2)Using K-means clustering
X193	16	10	X236	19	13
X260	18	10	X207	25	13
X195	21	10	X269	20	14
X234	21	10	X270	21	14
X222	22	10	X33	22	14
X232	22	10	X34	22	14
X146	23	10	X265	24	15
X134	24	10	X241	14	16
X135	24	10	X268	25	16
X173	26	10	X233	26	16
X198	17	11	X275	19	17
X93	18	11	X167	10	18
X271	20	11	X238	18	19
X267	23	11	X240	18	19
X38	25	11	X150	11	20
X159	26	11	X237	13	20
X257	23	12	X239	16	22
X256	24	12	X276	13	27
X171	27	12	X210	11	28
X50	15	13	X211	19	39
X52	15	13	X212	9	51

Table 1. shows the detected number of the quality factor for the two-step method of the statistical difference testing and k-means clustering. The name of the factor was expressed by replacing X1 to X292 according to order of the small steel bars manufacturing process.

- For Using statistical significance difference method: Among the 36 operation date, there are 29 operation date that have difference in the reprocessing ratio over 3. Because a maximum number of times derived by statistical difference testing for each factor is 29 times, when factor is derived over 9 times, it is supposed to quality factor. As a result of statistical difference testing, 100 factors were screened, leaving 192 factors.

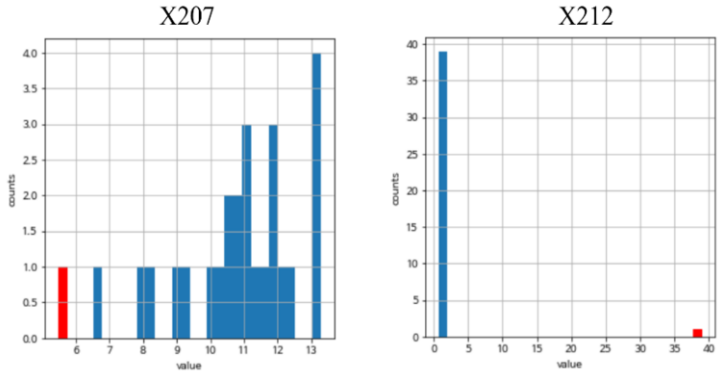

Figure 4. Example of X distribution detected as quality factor

- For Using K-means clustering(K=2) method: In 347 analyses according to the reprocessing rate value for each factor, each factor was derived from at least 0 times to up to 51 times. **Figure 4.** shows that distributions of the factors derived as the quality factor from a certain reprocessing ratio is divided into two groups through K-means clustering. After applying k-means clustering method, only 42 factors are remained. These 42 factors are identified as quality factors through a two-step method.

5. Conclusion

This study proposes novel method to overcome "the merged measurement problem" and "the variation due to the operation date": 1) the difficulty of applying general statistical analysis and data-mining techniques due to the characteristics of the steel process, 2) difficulty of applying general data preprocessing methods, 3) considering two realistic problems at the same time.

The first method, statistical difference testing, which verifies whether the mean difference between groups is statistically significant, provides the tendency of quality index for factor values from a macroscopic perspective. The second method, the method using K-means clustering(K=2), which compares the outlier ratio and reprocessing ratio for each factor, provides a microscopic perspective on the tendency of quality index when factor values have outlier. In this paper, through the two-step method, the macroscopic and microscopic perspectives of each method are harmoniously applied to the analysis for identifying quality factors.

In a situation where quality index are not labeled for each data, it is difficult to validate the results in identified quality factors. Therefore, it is difficult to verify the method proposed in this paper. However, as a result of applying the proposed method to the real small steel bar production data through the case study, identified quality factors include factors similar to the background knowledge of the field engineer. The quality factors of small steel bar obtained in this paper can be applied to the management of quality by searching for the optimal solution of quality factors with research for quality prediction.

However, both methods proposed in this paper have hyperparameters. In the statistical difference testing, the difference in reprocessing ratio and significant p-value between groups should be determined. In the method using K-means clustering, it is necessary to determine the acceptable range for the difference between the reprocessing ratio and the factor ratio. However, since it is hard to adopt supervised learning in this case study, it is difficult to evaluate the hyperparameter. Therefore, the decision should be made through the cooperation with the field engineer and meta-heuristic method. In addition, in the statistical difference testing, there is a problem that the mean difference between the two groups is sensitive to the number of samples. This problem was not considered because only under-bound was set in this study. The method using K-means clustering can detect only outliers present on one side of the distribution. If any factor responds to both outliers at both ends of the distribution, it will be difficult to search for these factors. It is expected that a more accurate search for quality factors will be possible if the following two points can be overcome in future research.

References

[1] D.C. Wang, H.M. Liu and J. Liu, “Research and Development Trend of Shape Control for Cold Rolling Strip,” Chin. J. Mech., 2017, 30, pp.1248-1261.

[2] G.W. Song, B. A. Tama, J. Park, J. Y. Hwang, J. Bang, S. J. Park and S. Lee, “Temperature Control Optimization in a Steel-Making Continuous Casting Process Using Multimodal Deep Learning Approach,” in Steel Res. Int., 2019, 90, 1900321.

[3] J. Jakubowski, P. Stanisz, S. Bobek and G. J. Nalepa, "Explainable anomaly detection for Hot-rolling industrial process," 2021 IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA), 2021, pp. 1-10.

[4] Y. Zhang and Y. Zhang, “Process monitoring, fault diagnosis and quality prediction methods based on the multivariate statistical techniques,” IETE Tech. Rev., Sep 2010, vol. 27, no. 5, pp. 406-420

[5] I. Mazur and T. Koinov, “Quality Control system for a hot-rolled metal surface,” Frattura ed Integrità Strutturale, 2016, vol. 10(37), pp. 287-296, doi:10.3221/IGF-ESIS.37.38.

[6] H.Y. Kim, H.C. Kwon, H.W. Lee, Y.T. Im, S.M. Byon and H.D. Park, “Processing map approach for surface defect prediction in the hot bar rolling,” Journal of Materials Processing Technology, 2008, Vol. 205, Issues 1–3, pp. 70-80.